
Open
MIND

Thomas Metzinger & Jennifer M. Windt (Eds). *Open MIND*.

Thomas Metzinger & Jennifer M. Windt (Eds).

Open MIND

Frankfurt am Main: MIND Group

*To our partners, Stefan Pitz and Anja Krug-Metzinger,
and to all the students and scholars of philosophy and cognitive science
who do not have easy access to scientific literature.*

Imprint

© 2015 by MIND Group, Frankfurt am Main

Philosophisches Seminar / Gutenberg Research College
Jakob Welder-Weg, 18
Johannes Gutenberg-Universität Mainz
D-55099 Mainz

Production: Satzweiss.com Print Web Software GmbH, Saarbrücken

ISBN: 978-3-95857-102-0

All rights reserved.

www.open-mind.net

Table of Contents

About this Collection – A Short Introduction to the Open MIND Project
Thomas Metzinger

General Introduction: What Does it Mean to Have an Open Mind?
Thomas Metzinger & Jennifer M. Windt

Target Papers, Commentaries, and Replies

- 1 Beyond Componential Constitution in the Brain: Starburst Amacrine Cells and Enabling Constraints**
Michael L. Anderson

Carving the Brain at its Joints - A Commentary on Michael L. Anderson
Axel Kohler

Functional Attributions and Functional Architecture - A Reply to Axel Kohler
Michael L. Anderson

- 2 What a Theory of Knowledge-How Should Explain – A Framework for Practical Knowledge beyond Intellectualism and Anti-Intellectualism**
Andreas Bartels & Mark May

The Semantic Reading of Propositionality and Its Relation to Cognitive-Representational Explanations - A Commentary on Andreas Bartels & Mark May
Ramiro Glauer

Preparing the Ground for an Empirical Theory of Knowing-How - A Reply to Ramiro Glauer
Andreas Bartels & Mark May

- 3 Introspective Insecurity**
Tim Bayne

“I just knew that!”: Intuitions as Scaffolded or Freestanding Judgements - A Commentary on Tim Bayne
Maximilian H. Engel

Introspection and Intuition - A Reply to Maximilian H. Engel
Tim Bayne

- 4 Meaning, Context, and Background**
Christian Beyer

Grasping Meaning - A Commentary on Christian Beyer
Anita Pacholik-Żuromska

Self-identification, Intersubjectivity, and the Background of Intentionality - A Reply to Anita Pacholik-Żuromska
Christian Beyer

- 5 The Puzzle of Perceptual Precision**
Ned Block

Phenomenal Precision and Some Possible Pitfalls - A Commentary on Ned Block
Sascha Benjamin Fink

Solely Generic Phenomenology – A Reply to Sascha Benjamin Fink
Ned Block

- 6 Rules: The Basis of Morality... ?**
Paul M. Churchland

Applied Metascience of Neuroethics - A Commentary on Paul M. Churchland
Hannes Boelsen

A Skeptical Note on Bibliometrics - A Reply to Hannes Boelsen
Paul M. Churchland

- 7 Embodied Prediction**
Andy Clark

Extending the Explanandum for Predictive Processing - A Commentary on Andy Clark
Michael Madary

Predicting Peace: The End of the Representation Wars - A Reply to Michael Madary
Andy Clark

8 Levels

Carl F. Craver

Mechanistic Emergence: Different Properties, Different Levels, Same Thing! A Commentary on Carl F. Craver
Denis C. Martin

Mechanisms and Emergence - A Reply to Denis C. Martin
Carl F. Craver

9 Mental States as Emergent Properties: From Walking to Consciousness

Holk Cruse & Malte Schilling

The “Bottom-Up” Approach to Mental Life - A Commentary on Holk Cruse & Malte Schilling
Aaron Gutknecht

The Bottom-Up Approach: Benefits and Limits - A Reply to Aaron Gutknecht
Holk Cruse & Malte Schilling

10 Why and How Does Consciousness Seem the Way it Seems?

Daniel C. Dennett

Qualia Explained Away - A Commentary on Daniel C. Dennett
David H. Baßler

How our Belief in Qualia Evolved, and Why We Care so much - A Reply to David H. Baßler
Daniel C. Dennett

11 The Heterogeneity of Experiential Imagination

Jérôme Dokic & Margherita Arcangeli

Imagination and Experience - A Commentary on Jérôme Dokic & Margherita Arcangeli
Anne-Sophie Brügger

The Importance of Being Neutral: More on the Phenomenology and Metaphysics of Imagination - A Reply to Anne-Sophie Brügger
Jérôme Dokic & Margherita Arcangeli

12 On the Eve of Artificial Minds

Chris Eliasmith

Future Games - A Commentary on Chris Eliasmith
Daniela Hill

Mind Games - A Reply to Daniela Hill
Chris Eliasmith

13 Can We Be Epigenetically Proactive?

Kathinka Evers

Should We Be Epigenetically Proactive? A Commentary on Kathinka Evers
Stephan Schleim

Understanding Epigenetic Proaction - A Reply to Stephan Schleim
Kathinka Evers

14 The Paradigmatic Body: Embodied Simulation, Intersubjectivity, the Bodily Self, and Language
Vittorio Gallese & Valentina Cuccio

Multisensory Spatial Mechanisms of the Bodily Self and Social Cognition - A Commentary on Vittorio Gallese & Valentina Cuccio
Christian Pfeiffer

Embodied Simulation: A Paradigm for the Constitution of Self and Others - A Reply to Christian Pfeiffer
Vittorio Gallese & Valentina Cuccio

15 All the Self We Need

Philip Gerrans

Memory for Prediction Error Minimization: From Depersonalization to the Delusion of Non-Existence - A Commentary on Philip Gerrans
Ying-Tung Lin

Metamisery and Bodily Inexistence - A Reply to Ying-Tung Lin
Philip Gerrans

16 Visual Adaptation to a Remapped Spectrum: Lessons for Enactive Theories of Color Perception and Constancy, the Effect of Color on Aesthetic Judgments, and the Memory Color Effect
Rick Grush, Liberty Jaswal, Justin Knoepfler & Amanda Brovold

What Can Sensorimotor Enactivism Learn from Studies on Phenomenal Adaptation in Atypical Perceptual Conditions? A Commentary on Rick Grush and Colleagues
Aleksandra Mroczko-Wąsowicz

Phenomenology, Methodology, and Advancing the Debate - A Reply to Aleksandra Mroczko-Wąsowicz
Rick Grush

17 An Information-Based Approach to Consciousness: Mental State Decoding
John-Dylan Haynes

What's up with Prefrontal Cortex? A Commentary on John-Dylan Haynes
Caspar M. Schwiedrzik

Can Synchronization Explain Representational Content? A Reply to Caspar M. Schwiedrzik
John-Dylan Haynes

18 Beyond Illusions: On the Limitations of Perceiving Relational Properties
Heiko Hecht

The Illusion of the Given and Its Role in Vision Research - A Commentary on Heiko Hecht
Axel Kohler

Manifest Illusions - A Reply to Axel Kohler
Heiko Hecht

19 The Neural Organ Explains the Mind
Jakob Hohwy

From Explanatory Ambition to Explanatory Power - A Commentary on Jakob Hohwy
Dominic L. Harkness

The Diversity of Bayesian Explanation - A Reply to Dominic L. Harkness
Jakob Hohwy

20 Millikan's Teleosemantics and Communicative Agency
Pierre Jacob

Communicative Agency and *ad hominem* Arguments in Social Epistemology - A Commentary on Pierre Jacob
Marius F. Jung

Assessing a Speaker's Reliability Falls Short of Providing an Argument - A Reply to Marius F. Jung
Pierre Jacob

21 Wild Systems Theory as a 21st Century Coherence Framework for Cognitive Science
J. Scott Jordan & Brian Day

Thickening Descriptions with Views from Pragmatism and Anthropology - A Commentary on J. Scott Jordan & Brian Day
Saskia K. Nagel

After Naturalism: Wild Systems Theory and the Turn To Holism - A Reply to Saskia K. Nagel
J. Scott Jordan & Brian Day

22 The Crack of Dawn: Perceptual Functions and Neural Mechanisms that Mark the Transition from Unconscious Processing to Conscious Vision
Victor Lamme

Consciousness as Inference in Time -
A Commentary on Victor Lamme
Lucia Melloni

**Predictive Coding Is Unconscious, so
that Consciousness Happens Now - A**
Reply to Lucia Melloni
Victor Lamme

**23 Vestibular Contributions to the Sense
of Body, Self, and Others**
Bigna Lenggenhager & Christophe Lopez

**Perspectival Structure and Vestibular
Processing - A Commentary on Bigna**
Lenggenhager & Christophe Lopez
Adrian Alsmith

**Vestibular Sense and Perspectival
Experience - A Reply to Adrian Alsmith**
Bigna Lenggenhager & Christophe Lopez

**24 Self-as-Subject and Experiential
Ownership**
Caleb Liang

**Are there Counterexamples to the
Immunity Principle? Some
Restrictions and Clarifications - A**
Commentary on Caleb Liang
Oliver Haug & Marius F. Jung

**Can Experiential Ownership Violate
the Immunity Principle? A Reply to**
Oliver Haug & Marius F. Jung
Caleb Liang

**25 Mathematical Cognition: A Case of
Enculturation**
Richard Menary

**Enriching the Notion of
Enculturation: Cognitive Integration,
Predictive Processing, and the Case of
Reading Acquisition - A Commentary on**
Richard Menary
Regina E. Fabry

**What? Now. Predictive Coding and
Enculturation - A Reply on Regina E. Fabry**
Richard Menary

**26 Understanding Others: The Person
Model Theory**
Albert Newen

**Multiplicity Needs Coherence –
Towards a Unifying Framework for
Social Understanding - A Commentary**
on Albert Newen
Lisa Quadt

**A Multiplicity View for Social
Cognition: Defending a Coherent
Framework - A Reply to Lisa Quadt**
Albert Newen

**27 Concept Pluralism, Direct Perception,
and the Fragility of Presence**
Alva Noë

The Fragile Nature of the Social Mind
- A Commentary on Alva Noë
Miriam Kyselo

Beyond Agency - A Reply to Miriam
Kyselo
Alva Noë

**28 How Does Mind Matter? Solving the
Content Causation Problem**
Gerard O'Brien

Does Resemblance Really Matter? A
Commentary on Gerard O'Brien
Anne-Kathrin Koch

Rehabilitating Resemblance Redux -
A Reply to Anne-Kathrin Koch
Gerard O'Brien

**29 Conscious Intentions: The Social
Creation Myth**
Elisabeth Pacherie

**Conscious Intentions: Do We Need a
Creation Myth? A Commentary on**
Elisabeth Pacherie
Andrea R. Dreßing

The Causal Role(s) of Intentions - A
Reply to Andrea R. Dreßing
Elisabeth Pacherie

30 Naturalizing Metaethics

Jesse Prinz

Conceptualizing Metaethics - A

Commentary on Prinz

Yann Wilhelm

**Should Metaethical Naturalists
Abandon *de dicto* Internalism and
Cognitivism? A Reply to Yann Wilhelm**

Jesse Prinz

**31 The Representational Structure of
Feelings**

Joëlle Proust

**The Extension of the Indicator-
Function of Feelings - A Commentary on**

Joëlle Proust

Iuliia Pliushch

Feelings as Evaluative Indicators - A

Reply to Iuliia Pliushch

Joëlle Proust

**32 The Avatars in the Machine:
Dreaming as a Simulation of Social
Reality**

*Antti Revonsuo, Jarno Tuominen & Katja
Valli*

**The Multifunctionality of Dreaming and
the Oblivious Avatar - A Commentary on**

Antti Revonsuo and Colleagues

Martin Dresler

**The Simulation Theories of Dreaming:
How to Make Theoretical Progress in
Dream Science - A Reply to Martin Dresler**

*Antti Revonsuo, Jarno Tuominen & Katja
Valli*

**33 Davidson on Believers: Can Non-
Linguistic Creatures Have
Propositional Attitudes?**

Adina Roskies

**Crediting Animals with the Ability to
Think: On the Role of Language in
Cognition - A Commentary on Adina Roskies**

Ulrike Pompe-Alama

Thought, Language, and Inner Speech

- A Reply to Ulrike Pompe-Alama

Adina Roskies

**34 Bridging the Objective/Subjective
Divide: Towards a Meta-Perspective
of Science and Experience**

Jonathan Schooler

**Bridging the Gap - A Commentary on
Jonathan Schooler**

Verena Gottschling

**Stepping Back and Adding
Perspective - Reply to Gottschling**

Jonathan Schooler

**35 The Cybernetic Bayesian Brain: From
Interoceptive Inference to
Sensorimotor Contingencies**

Anil K. Seth

**Perceptual Presence in the Kuhnian-
Popperian Bayesian Brain - A**

Commentary on Anil K. Seth

Wanja Wiese

Inference to the Best Prediction - A

Reply to Wanja Wiese

Anil K. Seth

**36 The Ongoing Search for the Neuronal
Correlate of Consciousness**

Wolf Singer

**It's Not Just About the Contents:
Searching for a Neural Correlate of a
State of Consciousness - A Commentary
on Wolf Singer**

Valdas Noreika

State or Content of Consciousness?

A Reply to Valdas Noreika

Wolf Singer

**37 Dreamless Sleep, the Embodied Mind,
and Consciousness**

Evan Thompson

Just in Time—Dreamless Sleep
Experience as Pure Subjective
Temporality - A Commentary on Evan
Thompson
Jennifer M. Windt

Steps Toward a Neurophenomenology
of Conscious Sleep – A Reply to Jennifer
M. Windt
Evan Thompson

38 What is the State-of-the-Art on Lucid
Dreaming? Recent Advances and
Questions for Future Research
Ursula Voss & Allan Hobson

Insight—What Is It, Exactly? A
Commentary on Ursula Voss & Allan
Hobson
Lana Kühle

Reflections on Insight - A Reply to Lana
Kühle
Ursula Voss

39 Representationalisms, Subjective
Character, and Self-Acquaintance
Kenneth Williford

Explaining Subjective Character:
Representation, Reflexivity, or
Integration? A Commentary on Kenneth
Williford
Tobias Schlicht

Individuation, Integration, and the
Phenomenological Subject - A Reply to
Tobias Schlicht
Kenneth Williford

About this Collection

A Short Introduction to the Open MIND Project

Thomas Metzinger

Author

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

1 What is this?

This is an edited collection of 39 original papers and as many commentaries and replies. The target papers and replies were written by senior members of the MIND Group, while all commentaries were written by junior group members. All papers and commentaries have undergone a rigorous process of anonymous peer review, during which the junior members of the MIND Group acted as reviewers. The final versions of all the target articles, commentaries and replies have undergone additional editorial review.

Besides offering a cross-section of ongoing, cutting-edge research in philosophy and cognitive science, this collection is also intended to be a free electronic resource for teaching. It therefore also contains a selection of online supporting materials, pointers to video and audio files and to additional free material supplied by the 92 authors represented in this volume. We will add more multimedia material, a searchable literature database, and tools to work with the online version in the future. All contributions to this collection are strictly open access. They can be downloaded, printed, and reproduced by anyone.

2 What is the MIND Group?

The MIND Group is an independent, international body of early-stage researchers, which I founded in 2003. It is formed of young philosophers and scientists with a strong interest in questions concerning the mind, consciousness, and

cognition. They come from various disciplines such as philosophy, psychology, cognitive science, and neuroscience.

Over the past decade, the MIND Group has cooperated with a number of institutions, such as the Frankfurt Institute for Advanced Studies, the *Meditationszentrum Beatenberg*, the *Wissenschaftskolleg zu Berlin*, and the *ICI Kulturlabor Berlin*. I first founded the group at the *Johannes Gutenberg-Universität* in Mainz in 2003, but soon had to relocate it to Frankfurt am Main, where we meet twice a year. Meetings typically involve two or three public lectures at the *Johann Wolfgang Goethe-Universität*, delivered by highly prominent guests, most of whom are now authors of the target papers in this collection and senior members of the group. In addition, our invited speakers offer extended, closed workshops, where advanced students have the opportunity to give short mock-lectures in English.

This format was inspired by a question which kept confronting me in my teaching: namely why are there so many excellent, smart young philosophers in Germany, who nevertheless are—and often remain—almost completely invisible on the international stage? More than half a century after World War II, only three or four German universities rank among the top 100. The established philosophical community is still largely disconnected from many of the latest and most exciting developments in modern philosophy of mind. One result of my thinking about this

problem was that this lack of integration into the global research context was caused, in part, by the language barrier. The biggest psychological obstacles for many young German philosophers seem to be, quite simply, to prepare a talk in English; find the courage to travel to an international conference in another country; and actually present their work there. One of the things we practice at MIND Group meetings is to prepare them for this.

The MIND Group sees itself as part of a larger process of exploring and developing new formats for promoting junior researchers in philosophy of mind and cognitive science. One of the basic ideas behind the formation of the group was to create a platform for people with one systematic focus in philosophy (typically analytic philosophy of mind or ethics) and another in empirical research (typically cognitive science or neuroscience). One of our aims has been to build an evolving network of researchers. By incorporating most recent empirical findings as well as sophisticated conceptual work, we seek to integrate these different approaches in order to foster the development of more advanced theories of the mind. One major purpose of the group is to help bridge the gap between the sciences and the humanities. This not only includes going beyond old-school analytic philosophy or pure armchair phenomenology by cultivating a new, type of interdisciplinarity, which is “dyed-in-the-wool” in a positive sense. It also involves experimenting with new *formats* for doing research, for example, by participating in silent meditation retreats and trying to combine a systematic, formal practice of investigating the structure of our own minds from the first-person perspective with proper scientific meetings, during which we discuss third-person criteria for ascribing mental states to a given type of system.

In addition to bridging geographical and disciplinary gaps, the MIND Group also aims to bridge conventional gaps produced by institutionalized hierarchies in academia. If you will, this is simply the academic variant of the generation gap: Few things are more intimidating to young researchers than being confronted, at a conference, with criticism from a researcher who has long been one of their intellectual heroes, known

only from textbooks, university classes, and research articles. For this reason, the MIND Group meetings have provided a protected space for promoting supportive and collegial interactions between senior and junior group members. In particular, the meetings of the MIND Group have helped establish and cement collaborations both among junior members and between junior and senior members. In some cases this has led to research visits, joint research projects, or long-term mentoring relationships. One motivation for founding the group, after all, was to smooth the path from university studies to being a professional academic for advanced students and young researchers.

3 Why did we do this?

We wanted to make a contribution by offering a freely available resource to others. When we first started thinking about what to do for the 20th meeting of the MIND Group, we knew we wanted it to be something special, some way of sharing with the interested academic public some of the expertise and collegial atmosphere we had built up over more than 10 years of working together. Initially we considered inviting everyone to a big four-day conference at an attractive location. But then we decided that we would do something more substantial and innovative - rather than creating a transient event and an enormous CO₂ footprint. We wanted to create a resource of lasting value that will subsist for years to come, and most importantly something that really is accessible for everybody—not only for people in affluent parts of the world, like ourselves. There seemed no better way to do this than by providing a large, open-access collected edition showcasing the work of our senior and junior members.

It quickly became clear that because of the scope of the project, and also because we had specific ideas about how it should be realized, this was going to be an experiment in autonomous open-access publishing. The MIND Group is an independent body, and apart from evening lectures by our invited speakers, its meetings are not open to the public. One goal of the Open MIND project was to first publish our scientific work without the support of a publisher, who would

eventually sell our own intellectual property back to us and our peers and simultaneously make it inaccessible to students in Brazil, India or China by locking it behind a paywall. We wanted to see if we could successfully establish a professional form of quality control via a systematic, journal-independent peer review process—and also if we could make it happen faster than existing and established institutions of academic publishing. We gave authors a deadline of 1st March 2014, and planned to publish the entire collection (including commentaries and replies) on January 15th 2015. We knew that these two pillars—speed and quality control—would be crucial to the success of the project. Academics are sometimes reluctant to publish their work in edited collections that often only appear years after the manuscripts have been submitted. We suspected that we would only succeed in obtaining state-of-the-art research papers if we could guarantee that the research discussed within them would not be out-of-date by the time the collection went online.

This publication format is also novel in another sense. Because a selected subset of junior group members acted as reviewers and commentators, the whole publication project is *itself* an attempt to develop a new format for promoting junior researchers, for developing their academic skills, and for creating a new type of interaction between senior and junior group members. Many of the reviewers and commentators in this edited volume have never actively participated in any scientific review process before, and, for many their commentary is their first ever publication. Throughout the project, all junior members were able to play different roles: they acted as reviewers, trying to improve and constructively criticize the target articles submitted by senior group members and commentaries submitted by their peers. Sometimes, reviewers were asked to go back and revise their reviews—and sometimes their reviews also led to the rejection of target papers altogether. They also acted as authors; and because their commentaries also went through a review process, they got to experience the review process from the other side as well.

This collection, therefore, is the result of a three-layered interaction between junior and senior members: personal (through meetings), ed-

itorial (through implementing a common publication project), and philosophical and scientific (through writing commentaries and replies). Throughout this process, we were often surprised and impressed by the results—and we hope that you will be, too.

4 Who did this?

Many people have made this contribution possible and many hours of unpaid work have gone into it. [Here](#) are the most important supporters.

4.1 The editors

As founder and director of the MIND group, I consider myself to be neither a junior nor a senior member. Therefore, I have not contributed a target paper or a commentary. If anything, my contribution lies in the choice and selection of authors and in the work, together with my collaborator Jennifer Windt, of bringing this project to completion.

4.2 Financial funding

All in all it has cost about € 241.000, to realize this project. First and foremost, the [Barbara Wengeler-Stiftung](#) needs to be mentioned: not only has it supported the current project with € 80,000, but over the years it has enabled the MIND Group to stay independent, and to realize a long series of fruitful meetings, during times when it was difficult to get support elsewhere. It has also supported some members by providing PhD and travel grants and by offering the annual € 10,000 *Barbara Wengeler-Prize*, awarded at our meetings in Frankfurt. The [Gutenberg Research College](#) and the [Volkswagen-Stiftung](#) have generously supported the project by providing two editorial staff positions for David Baßler, Daniela Hill, and Dr. Ying-Tung Lin, and by awarding a five-year Research Fellowship, beginning in April 2014, to me, Thomas Metzinger. This work was also partly supported by the European FP7 collaborative project [VERE](#) (contract no. 257695).

What Does it Mean to Have an Open MIND?

Thomas Metzinger & Jennifer M. Windt

Authors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Instead of an introduction

In our discussions leading up to the Open MIND collection's going online, we thought long and hard about how exactly to showcase the vast material in this collection and the ideas and motivations behind the project in our editors' introduction. We first thought about using the introduction to briefly summarize the take-home message of every single target article, commentary, and reply, as is customary in introductions to edited collections. This struck us, however, as being both unwieldy and redundant: it would have entailed summarizing and commenting on a total of 117 texts. More importantly, due to the online format of the collection (including in-text search functions) and the inclusion of abstracts and keywords in the papers themselves, the authors have already provided concise introductions to their own texts. Retracing their steps in an editorial introduction would not have added anything to the value and usability of the collection.

We then considered using the introduction to create our own personal best-of-Open-MIND list, discussing what we take to be the most valuable insights in every single article, or perhaps even focusing on the contributions that we personally take to be the most theoretically important. Though our own list of personal favor-

ites seemed to write itself naturally during the editing process, this strategy quickly struck us as being at odds with our motivation for creating the collection in the first place. Using the editors' introduction to create a personal best-of list would have been highly selective and biased by our own personal research interests and styles in a way that we felt would have contradicted our own ideal of open mindedness. In fact, for this reason, we decided to omit any references to the contributions to Open MIND in this introduction.

These considerations naturally gave rise to a more difficult and more profound question: What exactly do we mean by "open mindedness," not just in general, but in the context of interdisciplinary research on the mind? The strategy of using the contributions to the Open MIND collection as a foil for this more general academic variant of open mindedness was tempting. But again, we quickly realized that this approach would strike many readers (as well as, perhaps, some of our own authors) as highly idiosyncratic, arbitrary, or self-important.¹

¹ This is not, of course, to deny that we take "Open MINDedness" (as broadly practiced in the context of this collection) to be an example of "open mindedness" as a more general epistemic stance. And we are certainly proud enough of what we like to think of as our little star-collection to allow ourselves at least a few words on why we think this

So we decided to use our editors' introduction to briefly address a difficult, somewhat deeper, and in some ways more classical problem: that of what *genuine* open mindedness really is and how it can contribute to the Mind Sciences. The material in the collection speaks for itself. Here, and in contrast to the vast collection that is Open MIND, we want to be concise. We want to point to the broader context of a particular way of thinking about the mind. And we want to propose an account of what open mindedness could mean in the context of the contemporary, interdisciplinary Mind Sciences. This variant of open mindedness is characterized by epistemic humility, intellectual honesty, and a new culture of charity. It also has a pragmatic dimension: open mindedness of this kind is research generating and fosters an environment of sincere and constructive interdisciplinary collaboration. And it is profoundly inspired by the classical ideals of philosophy as a pursuit of genuine insight and rational inquiry, the importance of a critical and in a certain sense non-judgmental attitude, and the deep relationship between wisdom and skepticism as an epistemic practice. Finally, and again very classically, open mindedness has an ethical dimension

is the case. To begin with, many of the papers published here explore new ways of thinking, in the broadest sense, about the mind and new and innovative ways of driving research forward. In addition and perhaps most importantly, our choice of the title Open MIND reflects the idea that by introducing a two-way interaction between senior target authors and junior commentators through the review process, the commentaries and replies, we wanted to give our commentators the opportunity to enter into a discussion with more senior and prominent representatives of the field. Relatedly, the availability of the online version of the Open MIND collection to students and researchers from anywhere in the world, free of charge exemplifies theoretical and practical dimensions of what we consider to be academic open mindedness. And finally, on many levels, Open MIND was an exercise in editorial open mindedness. The authors and commentators asked to contribute to this collection were explicitly encouraged to discuss any topic they themselves thought relevant. The only restriction was that the target articles fall within the scope of the Mind Sciences. We also tried to foster a particular type of intellectual atmosphere by encouraging authors, commentators, and reviewers to be consistently constructive and charitable. Our hope was that this approach would bring out the best in our contributors in the different stages of the project. In many cases, we explicitly encouraged our authors to write in a way that would be accessible to readers from different academic backgrounds and to take different disciplinary perspectives into account. Generally, the publication of academic articles always involves a process of give and take between authors, editors, and reviewers. And we strongly felt that it would be a good indicator of the success of our collection if, at the end of the day, our authors were themselves happy and proud of their contributions. This entailed carefully calibrating our own roles as editors and in many cases leaving the final decision to our authors.

sion as well: it implies sensitivity to normative issues, including issues of an anthropological, sociocultural, and political kind. By bringing these different strands of ideas together and creating a bigger (and admittedly still sketchy) picture of what "open mindedness" might mean in the interdisciplinary Mind Sciences, we hope to start a conversation about how an open-minded attitude and a charitable culture of collaboration can be cultivated in the future. This is very much intended as an invitation to further think about and develop this topic. We hope our readers will join us in this endeavor.

2 Open mindedness as an epistemic stance

Open mindedness is not a theoretical position, but an epistemic practice. Clearly, there are many different kinds of open mindedness, and the precise way of characterizing the relevant kind will depend on the subject matter in question, or, more simply, on what it is that one is open minded about. As a first pass at a definition, we might say that open mindedness, in its most general sense, is characterized by epistemic humility and adherence to a general ideal of intellectual honesty. This is true for open mindedness in general, but also for the specific variants we are interested in here, namely open mindedness in academic research, including interdisciplinary scientific discourse on the mind.

Whatever else it may be, open mindedness is also an *attitude* that is now shared by a growing number of researchers in philosophy of mind, cognitive science, neuroscience, and artificial intelligence (AI). We are all interested in the deep structure of the human mind and of conscious experience, but we also recognize how far away we still are from a unified theoretical model that could satisfy philosophers and scientists alike, a model that is conceptually convincing, able to integrate all existing data and make use of different methods at the same time. We do not want to fool ourselves. Although great progress has been made during the last five decades, it is not at all clear which combination of methods and which type of theoretical approach will generate the final breakthrough

or even facilitate epistemic progress. We, meaning researchers of different stripes and from different disciplines comprising the Mind Sciences, including the authors contributing to this collection, are all in the same boat: we share a common epistemic goal, and we find ourselves working in a period of major historical transition. Progress in the empirical sciences of the human mind is certainly impressive and continuously gaining momentum, generating large amounts of new and sometimes surprising data. At the same time, exciting new approaches in formal modeling and philosophical meta-theory are increasingly opening up new perspectives. Yet it is not at all clear that we are already asking the right kinds of questions or exactly which combination of conceptual and empirical tools will do the trick. Seeing this fact clearly has already begun to change our attitude. Researchers from different disciplines are listening and talking to each other in new ways. Developing new forms of inter- (and intra-)disciplinary collaboration is an integral part of this process. “Having an open mind” also refers to a kind of scientific practice that involves honestly listening to representatives of exactly those approaches and academic disciplines that you may not have expected to make a contribution.

At the same time, open mindedness, understood as a fruitful and research-generating epistemic practice, should be clearly distinguished from arbitrariness, indecisiveness, lack of specificity, and, especially in the context of philosophy, lack of conceptual precision. Open mindedness is not just any kind of openness, and it is different from simply being non-committal or hedging. The challenge is to develop an understanding of open mindedness that is guided by theoretical considerations and empirical research findings alike. Ideally, this account should suggest specific strategies for cultivating forms of sincere interdisciplinary collaboration, sharpening the underlying conceptual issues, and developing precise predictions for future research. Open mindedness of this epistemically fruitful type will often be more about asking better questions than about committing to specific answers. It will involve an attitude of willingness to question or even reject one's own

prior commitments. It will be inherently critical (cf. [Lambie 2014](#)). And it will, perhaps, have more to do with striving for genuine understanding than with the search for truth and knowledge ([Taylor 2014](#)). One core idea of the great philosopher of science [Karl Popper](#), which is now reappearing in the latest mathematical theories of brain functioning, was that we are always in contact with reality at exactly the moment at which we falsify a hypothesis: the moment of failure is exactly the moment at which we touch the world.² Similarly, the best scientific theories will be those that most easily lend themselves to falsification. For this reason, open mindedness involves, among other things, endorsing very specific theoretical positions purely for the sake of epistemic progress, rather than for the sake of being right, advancing one's career, publishing in high-impact journals, and so on. Open mindedness is not so much about the specific content of a belief, be it personal or theoretical, but about the way in which it is held.

Searching for the right kinds of questions without considering the specific answers they are likely to generate or their immediate practical implications is a good first-order approximation to the specific type of attitude we are trying to describe. Another is to consider it as an interdisciplinary variant of the principle of charity. Our point is not just that philosophers should be empirically informed or that neuroscientists should listen carefully to constructive attempts at conceptual or methodological clarification. We need to develop a new culture of scientific investigation, and this will require new and sustainable forms of interdisciplinary collaboration. In philosophy, the “principle of charity” has long been recognized and pursued in the form of reading others' statements according to the best, strongest possible interpretation

² Here is what he said about the fundamental principle of any ideological form of rationalism turned *weltanschauung*: “Uncritical or comprehensive rationalism can be described as the attitude of the person who says ‘I am not prepared to accept anything that cannot be defended by means of argument or experience’. [...] Now it is easy to see that this principle of an uncritical rationalism is inconsistent; for since it cannot, in its turn, be supported by argument or by experience, it implies that it should itself be discarded” (cited from [Popper 2013](#), p. 435; originally in [Popper & Kieseewetter 1945/2003](#); see [Metzinger 2013c](#) for a popular discussion).

—that is, to never attribute irrationality, falsehoods, or fallacies to another if alternative and more charitable readings exist. But we also all know how hard this can be. Still, the point is to not gratuitously maximize disagreement with the aim of showcasing the novelty or importance of one's own arguments. Agreement should be optimized and as each other's interpreters, we should always, whenever possible, prefer the most coherent reading in order to maximize the truth or rationality of what another philosopher says. We now need an interdisciplinary variant of this principle, and not only in bridging the gap between the humanities and the so-called hard sciences of the mind, but also in organizing novel and more efficient forms of cooperation. This point applies not only to the relationship between disciplines, but also to that between different generations of researchers. An optimization problem has to be solved: What is the best way of pooling intellectual resources and of efficiently structuring research? Therefore, a second step toward approximating an undogmatic attitude of open mindedness is to characterize it as an openness to the possibility that, for mind and consciousness, there may be no such thing as a single leading or dominating discipline, no *Leitwissenschaft*, as we say in German. Rather, not only does the connectivity between already-existing research programs have to be strengthened, the overall pattern of scientific practice also requires a new internal structure. What is needed is a new and as we will argue genuinely philosophical way of thinking.

A genuine receptiveness to unexpected ideas and different disciplinary perspectives also presupposes a certain set of abilities and different types of epistemic virtues. Some of these may lie in the field of what is commonly, if somewhat vaguely, called “first-person methods”, for instance in the systematic cultivation of contemplative practice (i.e., the philosophically motivated development of *non-cognitive* and *non-intellectual* epistemic abilities). Another is tolerance of ambiguity: to not only tolerate transient cognitive, conceptual and theoretical inconsistencies between disciplines or generations, but to view certain kinds of ambiguity

as actually desirable, as a source of progress. Again, the challenge will be to distinguish productive types of ambiguity from those that are overly cautious or vague, hampering real progress. The same is true, of course, within academic disciplines themselves. Academic disciplines are not natural kinds. Contrary to what some might think, there may be no single authoritative or right way of doing philosophy, and there may be no clean way to distinguish philosophy from the empirical sciences. Open mindedness of the constructive kind will not waste time worrying too much about disciplinary demarcation criteria or labels, but will be open to different methods and approaches both between and within individual disciplines. Put differently, it may turn out to be less important whether a given question or position is philosophical (in the sense of being of a purely conceptual nature) or empirical than whether it genuinely helps advance the overall debate. Open mindedness clearly also has an inherently pragmatic dimension. When this kind of tolerance of ambiguity, for instance towards disciplinary borders, but also towards different (and ideally complementary) research methods is paired with conceptual clarity and precision, it becomes a driving force for research. This balancing act is what academic open mindedness is all about.

3 Open mindedness and the phenomenology of (un)certainty

Having an open mind involves, among other things, a specific way of being noncommittal with respect to the truth of a theoretical claim or proposition. As pointed out earlier, this is not the same as hedging: one can investigate and even defend the truth of a proposition or the adequacy of a given theoretical-conceptual or empirical model while at the same time acknowledging that it might be false. This continued openness to the falsifiability of scientific hypotheses, often associated with attempts to bring about specific ways of establishing and testing their falsity, is commonly regarded as a marker of good scientific practice. It is also the core of intellectual honesty. As [Russell](#) tells us,

“intellectual integrity [is] the habit of deciding vexed questions in accordance with the evidence, or of leaving them undecided where the evidence is inconclusive” (2009, p. 579). The moment at which we give up this openness is the moment at which we lapse into dogmatism. The real danger, says Russell, is never the content of a doctrine, be it religious or political, but always “the way in which the doctrine is held” (Russell 2009, p. 582). Of course, this intrinsic connection between wisdom and not-knowing has long been recognized (Ryan 2014). In the *Gorgias*, Socrates explicitly claims that he is happy to be refuted if he is wrong. In fact, he claims he would rather be refuted than to refute someone else because it is better to be delivered from harm oneself than to deliver someone else from harm. And in the *Apology* (21d), after being accused of blasphemy and of corrupting the youths of Athens, Socrates famously states, before the tribunal of 501 Athenians, “I neither know nor think that I know”. Both in Western and in Eastern philosophy, the acknowledgment of not-knowing has long been regarded as an antidote to epistemic harm.

This is not the place to enter into a discussion of open mindedness in the context of the philosophy of science or to trace the history of philosophical theorizing about the concept of “wisdom”. We do, however, want to draw attention an important point: open mindedness as an epistemic practice involves a specific kind of mental attitude and is closely related to certain kinds of phenomenal states. Cultivating the relevant kinds of conscious states and epistemic attitudes makes a real difference, or so we suspect, by facilitating the development of a research climate that is conducive to constructive and genuinely fruitful discourse and new forms of collaboration. This is an empirical prediction, and it could turn out to be false. For now, our claim is that the kind of open mindedness we describe here is needed if we are even to begin investigating the truth of this prediction. If, at the end of the day, this strategy should fail — that is, if there turn out to be good empirical reasons for rejecting the claim that there actually are specific phenomenological profiles and mental attitudes that decisively facilitate pro-

gress in interdisciplinary research on the mind—this would be a valuable insight. But this insight about the value of open mindedness in scientific discourse itself depends on an initial willingness to cultivate exactly the kind of epistemic practice in question.

If this is right, there is another reason to be interested in open mindedness in the present context. This is that open mindedness, as an epistemic practice and mental attitude, is itself a potential target for interdisciplinary consciousness research. Philosophy of mind in particular can contribute by laying the theoretical–conceptual groundwork for the further empirical investigation of open mindedness in academic life and proposing points of contact with psychology and cognitive neuroscience. To make this inner connection more clearly visible, we will now briefly sketch the outlines of such an account.

Where might one begin investigating open mindedness as a mental state? At the outset, it stands to reason that the relevant form of open mindedness has precursors in the history of philosophy and might also be interestingly related to current debates on philosophical methodology. After all, the principles of epistemic humility, intellectual honesty, charitability, and searching for more accurate questions while cultivating a productive form of tolerance of ambiguity are deeply rooted in the history of philosophy. On a systematic and more general level, one would expect philosophy, as the discipline traditionally most concerned with the status of knowledge and truth and the practice of inquiry itself, to be able contribute to an analysis of what open mindedness really is. Based on these considerations, four questions seem particularly relevant: one, what is the relationship between open mindedness, intuitions, and philosophical methodology? Two, what is the relationship between open mindedness and the tradition of philosophical skepticism? Three, what would answers to the first two questions tell us about the relationship between open mindedness and the allegedly most pressing problem for interdisciplinary consciousness research, the subjectivity of phenomenal mental states? Might we even use the analysis of open mindedness to formu-

late principles for the investigation of phenomenal states and the status of first-person data? And four, how is open mindedness as an epistemic stance related to ethical and practical questions? For instance, how can the analysis of open mindedness contribute to normative issues related to neurotechnological interventions in the human brain? And does it lead to any specific suggestions on how to cultivate new forms of interdisciplinarity?

3.1 Intuitions and the phenomenology of certainty

The concept of intuition has a long philosophical history and is also firmly rooted in everyday language and folk psychology.³ Intuition, in everyday language, refers to immediate and direct insight, independent of reflection, to instinctively grasping or sensing a matter of fact. In the history of philosophy, the concept of intuition often has dual epistemic and experiential readings, and this is true for the traditions of rationalism and empiricism alike. In the *Rules for the Direction of the Mind* (Rule 3), Descartes describes intuitions as an immediate, effortless, and indubitable kind of seeing with the mind, which is even more reliable than deduction. In his *Essay Concerning Human Understanding* (IV.II.I), Locke tells us that intuition involves a direct perception of ideas that is, once more, the basis of all forms of knowledge. The close relationship between intuitions and sensory perception, and especially seeing, is already evident in the Latin verb *intueri*, which means to look and observe, but also to examine or consider. The central underlying element is the immediacy and directness of perception, which is imported into the concept of intuition via an implicit analogy between the phenomenology of sensory perception and genuine insight in an epistemic sense.

The epistemic status of intuitions, as well as different ways of defining the concept of intuition, are a matter of controversy in the current debate on philosophical methodology. The debate on intuitions stands at the center of the

confrontation between classical and allegedly intuition-based conceptual analysis conducted in the proverbial philosophical armchair (for critical discussion, see Cappelen 2012) and recent claims from experimental philosophy. Experimental philosophy typically involves collecting laypersons' responses to vignettes inspired by well-known philosophical thought experiments (for discussion, see Knobe & Nichols 2008; Alexander 2012; for a general introduction to intuitions in philosophy, see Pust 2014). These questionnaires are supposed to offer a new, empirically-based method for investigating intuitions and the underlying cognitive mechanisms. According to some experimental philosophers (for discussion and further references, see Alexander & Weinberg 2007), the results of these types of studies cast doubt on the reliability of intuitions as a mark of philosophical expertise. Intuitions, in this view, are simply too variable and context-dependent to count as insights in any deep, epistemologically interesting sense.

Here, we would like to propose a definition of intuitions that is compatible with the historical literature as well as being phenomenologically and empirically plausible. Departing from our brief remarks on the history of intuitions in philosophy, we suggest that intuitions are the “phenomenal signature of knowing”, a seemingly direct and effortless way of perceiving or seeing with one's mind arising independently of a prior process of reflection. The analogy between intuiting and perceiving provides an entry point for a naturalized concept of intuition. But it also suggests a potentially dangerous equivocation between phenomenological and epistemological readings of the concept of intuition. If the phenomenology of intuiting is indeed similar to that of perceiving in virtue of its effortless and seemingly direct experiential quality, then this immediately poses the problem that the phenomenology of intuiting and perceiving can be deceptive: what seems, subjectively, to be a case of veridical perception can always turn out to be a hallucination or an illusion (for an introduction to the problem of perception, see Crane 2014), or a nocturnal dream (see Windt & Metzinger 2007; Metzinger 2013a; Windt 2015). Similarly, what seems to bear the

³ This section draws on arguments first presented in Metzinger & Windt (2014).

marks of genuine insight can always turn out to be an epistemic illusion.⁴

If intuitions are indeed mental states characterized by a specific phenomenology, this suggests that the attempt to simultaneously characterize them both as involving genuine insight and as the basis of knowledge rests on what elsewhere we call the “*E-error*”: a category mistake in which epistemic properties are ascribed to something that does not intrinsically possess them (Metzinger & Windt 2014, p. 287). If our account of intuitions is on the right track, then intuitions are potentially dangerous, because in virtue of their phenomenology and their possessing an occurrent conscious character of “insight”, they predispose us to believe certain propositions merely on the basis of seemingly “understanding” them. The phenomenology of intuitions is such that it immediately and effortlessly creates a bias towards accepting the truth of propositions that, subjectively, we simply *know* or feel to be true, while simultaneously preventing us from seeking further justification, because these truths also seem unconstructed, indubitable, and self-evident. In this view, one of the factors underlying intuitions and intuitive

plausibility is that, because of their phenomenal character, they prevent open-minded inquiry. Intuitions turn us into inner dogmatists. And this is true not only for individual propositions held to be intuitively true, but also for continued adherence to theoretical claims about the status of intuitions as a guide to or even as the basis of knowledge and genuine insight. The phenomenal character of intuitions even predisposes us towards certain meta-theoretical intuitions about the general epistemic status of intuitions, and we can see the marks of this throughout the history of philosophy as well as in contemporary debate (e.g., Bealer 1998; Chudnoff 2013). The analysis of intuition clearly should not itself be driven by intuitions. Instead, this is a prime example of where an open mind is needed.

Our own account starts out from the assumption that intuitions are a specific class of phenomenal states. Human beings can direct their introspective attention toward the content of the relevant states and, at least partly and under certain conditions, report on it. Many higher animals very likely also possess intuitions even if they are not able to directly attend to or verbally report on their intentional contents. Before the evolution of biological nervous systems and before the emergence of phenomenal consciousness, no intuitions existed on our planet. Patients in coma or human beings in unconscious, dreamless sleep have no intuitions in the sense intended here. At the same time, intuitions probably have a long evolutionary history: there must have been a point in time at which the first intuition appeared in the mind of some conscious organism and this specific type of inner state then propagated itself across thousands of generations while its functional profile became ever more differentiated. Plausibly, one could describe the having of intuitions as an *ability*—a mental ability that was adaptive and that was acquired gradually.

If one takes the phenomenal character of intuitions seriously, this ability clearly seems to be an epistemic ability: *prima facie*, to have an intuition means to have the subjective experience of knowing something, directly and immediately, without necessarily being able to ex-

⁴ For a striking case study of two patients who experienced strong feelings of subjective certainty, including religious beliefs, during epileptic seizures, see Picard (2013). These cases are particularly interesting as these beliefs seemed entirely convincing during the seizures, even though they contradicted the patients’ longstanding convictions. It is interesting to see the connection to what earlier, we called the “ability to tolerate ambiguity”: While conceptually, “certainty” involves “knowing that one knows” (or *maximal epistemic precision*), on a purely formal level describing the underlying brain dynamics, epistemic precision is the inverse of variability, or the “confidence” the system places in a source of sensory information about the external world (Picard & Friston 2014). Empirical research suggests that it is the functional role of the anterior insula to signal uncertainty, the fact “that there is something we do not understand” (Picard 2013, p. 2497). The representation of uncertainty and ambiguity, in turn, causes an aversive affective state, often involving feelings of discomfort and anxiety of the type we continuously try to minimize. By contrast, direct electrical stimulation of a small area in the anterior-dorsal insula causes intense feelings of bliss (Picard et al. 2013), and it has been suggested that such blissful states, if occurring in the context of epileptic seizures, are associated with maximized coherence of the phenomenal self-model (PSM; Metzinger 2003). Subjectively, this coherence is expressed by a dramatically heightened sense of self, by an intense phenomenal experience of presence, integratedness, harmony with the world, plus intense positive affect (for five case reports, see Picard & Craig 2009). For human beings, ambiguity is not easy to tolerate, because it presents a constant threat to the coherence of our PSM, and cultivating such tolerance requires developing the functional ability to de-identify from the aversive affective states and the “epistemic anxiety” that automatically accompanies them. Tolerance of ambiguity, it seems, demands courage and a specific form of choiceless awareness.

press this knowledge linguistically or to provide an epistemic justification. Typically, inner experience seems to present knowledge to the subject of experience, even if one does not know *how* and *why* one possesses this knowledge. Intuitions are the phenomenal signature of knowing, a seemingly direct form of “seeing” the truth. As soon as we ascribe epistemic status to intuitions on the basis of their phenomenology alone, however, we commit the E-error. “Epistemicity”, the phenomenal quality of “insight” and “comprehension”, or the feeling of being a knowing self, as such is only a phenomenal quality, just as redness, greenness, and sweetness are. One well-known philosophical problem is that the phenomenological and epistemological readings can always come apart, because what phenomenologically appears as a kind of perception could really be a hallucination or an illusion. Subjectively indistinguishable mental states do not necessarily have the same epistemic status. Trivially, the difference between veridical perception and hallucination (in the philosophical sense; see [Macpherson 2013](#); [Crane 2014](#)) is not available on the level of subjective experience itself, and therefore the confusion between phenomenal character and epistemic content is naturally grounded in the transparent phenomenology, the seeming directness and immediacy of the relevant kinds of phenomenal states. The same is true for the phenomenology of intuition. Conflating epistemic status and phenomenal character becomes particularly dangerous if it is imported into theoretical debates, and if the phenomenal quality in question is that of “epistemicity”, of direct and non-inferential knowing itself. The important lesson is that *as* phenomenal states, such states are neither necessarily veridical nor necessarily non-veridical. Experience as such is not knowledge. *As* subjective experiences, these states possess no intentional properties and cannot be semantically evaluated by concepts like “truth” or “reference”. Phenomenal transparency is not epistemic transparency.

Many, but not all, of our philosophically relevant intuitions are characterized by an additional element of *certainty*, of *just knowing* that one knows. Here, the phenomenal signature of

knowing does not only refer to the content of what is seemingly known in a direct, and non-inferential manner, but to our higher-order, subjectively-experienced knowledge itself. This means that the phenomenal character of “epistemicity” that accompanies and tags the respective mental content as an instance of knowing has itself become transparent. Its representational character is not introspectively available anymore: the fact that epistemicity is itself the content of a non-conceptual mental representation, that it is internally constructed and always contains the possibility of misrepresentation, is veiled by an experience of immediacy. Transparency is a special form of darkness. Something constructed is experienced as a *datum*, as something given. Therefore, in stable intuition states we not only experience the first-order content as directly given, but the epistemicity of the state itself. Let us call such states *intuitions of certainty*. Referring to [G. E. Moore](#)⁵ one might say that the phenomenal signature of knowing has itself become diaphanous or transparent: according to my own subjective experience, I simply *know* that I know, and the possibility of error and falsehood is not given on the level of conscious experience itself. From the fact that a conscious perception instantiates the phenomenal quality of “greenness” it does not follow that the underlying process or even the perceptual object are green. The same is true for the “phenomenal signature of knowing” that characterizes intuitions.

Intuitiveness is a property of theoretical claims or arguments, relative to a class of representational systems exhibiting a specific functional architecture. Conscious human beings are one example of such a class. The brains of human beings are naturally evolved information-processing systems, and when engaging in explicit, high-level cognition they use specific representational formats and employ characteristic

5 In *The Refutation of Idealism*, [G. E. Moore](#) wrote: “The term ‘blue’ is easy enough to distinguish, but the other element which I have called ‘consciousness’—that which a sensation of blue has in common with a sensation of green—is extremely difficult to fix. [...] And in general, that which makes the sensation of blue a mental fact seems to escape us; it seems, if I may use a metaphor, to be transparent—we look through it and see nothing but the blue; we may be convinced that there is something, but what it is no philosopher, I think, has yet clearly recognized” (1903, p. 446).

styles of processing. Whenever we try to comprehend a certain theory, an argument or a specific philosophical claim, our brains construct an internal model of this theory, argument, or claim (Johnson-Laird 1983, 2008; Knauff 2009). This mostly automatic process of constructing mental models of theories possesses a phenomenology of its own: some theories just “feel right” because they elicit subtle visceral and emotional responses, some claims “come easily”, they are experienced as sound and healthy, and some arguments (including the implicit assumptions upon which they rely) seem “just plain natural”. Some forms of skepticism appear “healthy” to us, while others do not—there seems to be a deep connection between sanity and reason.

There may be two overarching reasons for this well-known fact. First, theories that are intuitively plausible exhibit a high degree of “goodness of fit” in regard to our network of explicit prior convictions. More generally, they optimally satisfy the constraints provided by our conscious and unconscious models of reality as a whole. These microfunctional constraints implicitly represent both the totality of the knowledge we have acquired during our lifetime and certain assumptions about the deep causal structure of the world that proved functionally adequate for our biological ancestors. Theories that immediately feel good because they are characterized by a high degree of intuitiveness maximize a specific kind of internal harmony. What we introspectively detect is a high degree of consistency, but in a non-linguistic, subsymbolic medium. Therefore we could also replace the term “intuitiveness” with a notion like “intuitive soundness” or “introspectively detected consistency or goodness of fit” (relative to a preexisting model of reality). In principle it should be possible to spell out this point on a mathematical level, by describing the underlying neural computations and their properties in a connectionist framework, or by utilizing the conceptual tools provided by dynamical systems theory or predictive coding.

A second perspective might be to look at intuitions not from a representationalist, but from biophysical perspective. We are embodied

beings, and there are different levels of embodiment (Metzinger 2014). Computational, but also thermodynamical imperatives guide the self-organization of representational states in our brains. One major causal factor underlying the conscious experience of “intuitive soundness” might simply be the amount of energy it takes to activate and sustain a mental model of a given theory, plus the amount of energy it would take to permanently *integrate* this theory into our pre-existing model of reality. Our mental space of intuitive plausibility can in principle be described as an energy landscape: claims that “come easily” do so because they allow us to reach a stable state quickly and easily, theories that “feel good” are theories that can be appropriated without a high demand of energy. Theories that *don’t* feel good have the opposite characteristics: they “don’t add up”, they “just don’t compute”, because they endanger our internal harmony and functional coherence, and it would take a lot of energy to permanently integrate them into our overall mental model of reality. They are costly. In a biophysical system like the human brain there may well be a direct connection between thermodynamic efficiency and reduction of complexity on the level of information processing. If biological self-organization involves continuously minimizing the prediction errors generated by the flow of “hypotheses” originating in the brain’s current model of reality, then the process that creates what today we call our deepest “theoretical intuitions” may also be described as such an attempt to reduce variational free energy. While on a more abstract level this process can be said to minimize representational complexity while simultaneously maximizing the evidence for the overall model, it is also a physical process that is not guided by abstract rationality constraints, but simply one that optimizes metabolic and statistical efficiency at the same time (Sengupta et al. 2013; Friston 2010; Hohwy 2013).

We need an open mind, because many of the best future theories about the human mind and conscious experience may just “not compute” for beings like us. However, what does or does not compute is, in part, a contingent fact determined by the functional architecture of our

brain, shaped by millions of years of biological evolution on this planet, as well as—to a much lesser degree—by our individual cognitive history and a given cultural/linguistic context. The phenomenology of intuitive soundness—the fact that some arguments seem “just natural”—is a biological phenomenon that is additionally supported by a short cultural history of cognitive niche construction. In this framework, the space of intuitive plausibility reflects exactly those aspects of our evolutionary history and of our more recent cognitive niche that have become transparent—that we have long ceased to experience as evolved and culturally driven, but regard as unconstructed, immediate, and even indubitable. Importantly, the inner landscape of our space of intuitive plausibility is not simply contingent on our evolutionary history and on certain physical and functional properties of our brains—it was optimized for *functional adequacy* only. This process of optimization serves to maximize reproductive success and to sustain an organism’s coherence and physical existence, but this does not mean that the *content* of intuitions is epistemically justified in any way. This is especially true because the evolved functional adequacy of intuitions applies to everyday action in practical contexts and ancestral environments—not to abstract reflection in theoretical contexts or cognitive environments. This is why searching for a comprehensive theory of the conscious mind presents such a major challenge to our intellectual honesty: it demands that we investigate a claim even if it contradicts our deepest intuitions, even if it cries out for a more moderate, weaker version because it just “doesn’t compute” and somehow seems “just too radical”, costly, painful or even self-damaging. In this view, any philosophical methodology that just tries to make our “deepest intuitions” explicit in a conceptually coherent manner appears to be a rather trivial enterprise. If our claims here are correct, then intuition-mongering may even border on intellectual dishonesty. At best, it just charts our intuition space; at worst, it confuses failures of imagination with insights into conceptual necessity (“philosopher’s syndrome”, according to [Dennett 1991](#), p. 401).

3.2 Suspending judgment, inner quietude, and the phenomenology of uncertainty

If intuitions can be described as creating a transparent inner bias and perhaps even as involving an inner form of dogmatism, then we might, it would seem, make progress in understanding open mindedness as a mental state by looking to cases characterized by the phenomenal signature of *not* knowing and of uncertainty. The philosophical tradition of skepticism seems to be a promising place to look. Skepticism comes in many different strengths and flavors (see [Landesman 2002](#) for a comprehensive introduction), but what is distinctive about philosophical skepticism is perhaps best captured by the meaning of the original Greek term, where skeptic (related to the Greek verb *sképtomai*) refers, quite simply, “to one who inquires into the truth of things or wishes to gain knowledge about some subject matter” ([Landesman & Meeks 2003](#), p. 1). Skeptical inquiry, in the philosophical sense, is not so much concerned with the truth of particular beliefs or theoretical claims as with the possibility of knowledge and certainty in a more fundamental sense. It also does not always aim at denying the truth of our most basic beliefs by construing outlandish skeptical hypotheses such as the Cartesian evil genius. Generally, skeptical arguments cast doubt on commonly (and often implicitly and unreflectively) accepted means for attaining knowledge—and in so doing frequently give rise to new and fruitful discussions on how our epistemic practices might be improved. Throughout the history of philosophy, skepticism, at its best, has often been deeply constructive and has enabled genuine progress.

The philosophical tradition that has perhaps been most concerned with cultivating a skeptical attitude and with uncertainty and not-knowing as a mental state, at least in Western philosophy, is Pyrrhonian skepticism, which was one of the two major schools of skepticism in antiquity. Here, we want to tentatively suggest that it could be instructive to trace many of the aspects that we claim characterize open mindedness all the way back to the Pyrrhonian skeptics. This claim might strike some as sur-

prising, because Pyrrhonian skepticism is often seen as a particularly radical and excessive kind of skepticism (Hume's *Enquiry Concerning Human Understanding* is a classical example of this). It is fair to say that in contemporary philosophy, Pyrrhonian skeptics are an endangered species (for an introduction, see Fogelin 1994; Sinnott-Armstrong 2004; especially Stroud 2004; Fogelin 2004), with the tradition often being regarded as a bit of a historical oddity. This is fueled by what little is known of its founding father, Pyrrho of Elis (c. 360 to c. 270 BCE). Most of this is anecdotal, as Pyrrho wrote nothing himself (Bett 2014). Diogenes, for instance, tells us that Pyrrho:

led a life consistent with this doctrine, going out of his way for nothing, taking no precaution, but facing all risks as they came, whether carts, precipices, dogs, or what not, and, generally, leaving nothing to the arbitrage of his sense; but he was kept out of harm's way by his friends, who [...] used to follow close after him. (1943, 9.62)

Pyrrho did not return the favor, reportedly passing by an acquaintance who had fallen into a slough without offering him any help (*ibid.*, 9.63). Clearly, this is a far cry from the constructive and research-generating type of open mindedness we hope to promote here.

A more thoughtful and differentiated account can be found in Sextus Empiricus's (1987) treatment of skepticism, where he refers solely to Pyrrhonian skepticism.⁶ According to Sextus:

Skepticism is an ability, or mental attitude, which opposes appearances to judgments in any way whatsoever, with the result that, owing to the equipollence of the objects and reasons thus opposed, we are brought firstly to a state of mental suspense and next to a state of 'unperturbedness' or quietude. (1987, Chapter 4)

⁶ Sextus distinguishes three types of philosophers by their adherence to different types of systems: dogmatists, or those who claim to have discovered the truth; academics, who deny that the truth can be apprehended; and skeptics, who continue to inquire.

Clearly, there is at least a superficial similarity between Sextus's claim that skepticism is an ability and our description of open mindedness as an epistemic practice. Here, we briefly review the most important characteristics of Pyrrhonian skepticism and argue that there indeed exist a number of insightful parallels to open mindedness as an epistemic practice.

A first point is that from the perspective of Pyrrhonian skepticism, dogmatism is the end of reasoning and the opposite of philosophical reflection. At the same time, the anti-dogmatism of the Pyrrhonian skeptics did not prevent them from giving "assent to the feelings which are the necessary results of sense-impressions" (1987, 7.13). The Pyrrhonian skeptics merely withheld assent to "the non-evident objects of scientific inquiry" (*ibid.*, 7.13). As an early form of what we call academic open mindedness, Pyrrhonian skepticism was directed, first and foremost, "against the dogmas of 'Professors'—not the beliefs of common people pursuing the honest (or, for that matter, not so honest) business of daily life. The Pyrrhonian skeptic leaves common beliefs, unpretentiously held, alone." (Fogelin 2004, p. 163)

This suggests that if we want to contrast the cultivation of an anti-dogmatic mindset with intuitions, this point should be applied not to intuitions and feelings of certainty in general, but to philosophical intuitions in particular. Philosophical intuitions, in virtue of their distinctive phenomenal character, involve a specific and often highly-specialized form of inner dogmatism: they quickly and effortlessly create an inner bias towards a given theoretical position, while at the same time making it seem so indubitable and certain as to prevent further critical inquiry. Even though the terminology is, of course, different, the Pyrrhonian attitude of anti-dogmatism presents itself as an antidote to exactly the type of uncritical, judgmental attitude that is the hallmark of intuitions.

Second, the Pyrrhonian skeptic, in his quest for "quietude in respect of matters of opinion and moderate feeling in respect of things unavoidable" (Sextus 1987, 12.25), makes use of stereotyped tropes or modes of argument. The tropes are all very similar in structure, in-

volving a series of contrasts between opposing statements, with the aim of leading to irresolvable disagreement and inducing a suspension of judgment. True to the characterization of the Pyrrhonian skeptic as one who inquires, “the modes [...] were not designed to inhibit reasoning. Rather, they were designed to assist the Pyrrhonian in continuing to inquire by shielding her from the disquieting state of dogmatism” (Klein 2014). As Sextus (1987, 7.13) tells us, the Pyrrhonian, when entering into a debate with the dogmatist, does not assert his arguments in the manner of claiming their truth; instead, he asserts them only provisionally and purely for the sake of argument, enabling him to practice epoché, or to bracket his assumptions about the truth of the relevant propositions. The tropes, then, are not just a strategy for convincing one’s opponent, but a specific way of cultivating this more general kind of epistemic attitude:

Like piano exercises for the fingers that would result in semi-automatic responses to the printed notes on a sheet of music, the modes were mental exercises that would result in semi-automatic responses to claims being made by the dogmatists—those who assented to the non-evident. (Klein 2014)

We certainly do not mean to suggest that we should all become Pyrrhonian skeptics by formulating modernized versions of the tropes. We only want to point out that the naturalistic strategy of preparing and then handing over questions to scientific research can be viewed as fulfilling a similar function, namely as cultivating the epistemic virtues and abilities associated with open mindedness. This in itself, of course, is nothing new. A similar idea can be found, for example, in Russell’s claim that,

as soon as definite knowledge concerning any subject becomes possible, this subject ceases to be called philosophy, and becomes a separate science. [...] those questions which are already capable of definite answers are placed in the sciences, while

those only to which, at present, no definite answer can be given, remain to form the residue which is called philosophy. (Russell 1912/1999, p. 112)

Following Russell, philosophy itself is a specific variant of cultivating what, earlier, we called a tolerance of ambiguity, and its value is “to be sought largely in its very uncertainty” (1912, 113). The Pyrrhonian tropes are just one example from the history of philosophy of how a particular style of argumentation can be used not just to generate particular insights but also to promote a particular style of thinking. Analogously, one of the reasons why interdisciplinary collaboration and data-driven arguments in philosophy are valuable may be that they are a way of practicing and cultivating open mindedness. Interdisciplinary research projects don’t just produce new data, but leave their marks on the minds of the researchers involved as well.

Third, the suspension of judgment, which is the outcome and in some sense the aim of the modes, is described by Sextus as a state of mental rest and as an “untroubled and tranquil condition of the soul” (1987, 4. 10). It also, however, has a normative dimension, involving the claim that if there is irresolvable disagreement between two opposing positions, one should refrain from adopting either of them.

In the ambiguity between these two readings, there is a nice point of contact between open mindedness as a mental state and something that today one might call the ethics of belief (Clifford 1877/1999; Chignell 2010) and of belief formation. There is clearly a social (Goldman 2010) and perhaps even an interdisciplinary dimension of epistemology, both in a theoretical and in practical sense. As is the case for the dialectical confrontation between the Pyrrhonian and the dogmatist, progress (in the sense of suspension of judgment) will often result from confronting one’s own convictions with those held by others, as well as from confronting them with real-world counterexamples.⁷ By contrast, accumulating evidence suggests that

⁷ This reliance on actual cases of disagreement, rather than on hypothetical scenarios and thought experiments, is also one of the differences between Pyrrhonian and Cartesian skepticism.

merely simulating this process by charting one's own intuitive responses to carefully calibrated thought experiments is not nearly as effective, and is actually often quite misleading (Gendler & Hawthorne 2010; Alexander 2012; Dennett 2013). Doing, as the Pyrrhonian skeptics realized, is better than merely imagining.

Indeed, empirical evidence suggests that our natural confidence in naïve realism is so strong that it remains largely unscathed by theoretical evidence to the contrary. In one study, when participants read a text about cognitive limitations and biases, this did not affect their confidence in their own social judgments. Confidence levels were only significantly reduced when theoretical challenges to naïve realism were presented alongside specific examples, such as visual illusions. As the authors put it, “acknowledging susceptibility to bias [...] may not always translate to actually tempering one's confidence or expressing an openness to change. Instead, experiencing unconscious cognition and bias was required to reduce confidence and closed-mindedness” (Hart et al. 2015, 6). This acknowledgment of the value of the practical and experiential dimensions of suspending judgment is implicit in the Pyrrhonian tropes.

Fourth, ataraxia, or quietude, according to Sextus, automatically and effortlessly follows on the heels of the suspense of judgment. This unintentional character of quietude is important, because it means not only that quietude cannot be actively brought about, but also that it is found in a place quite different from that in which one was looking:

the Skeptics were in hopes of gaining quietude by means of a decision regarding the disparity of the objects of sense and of thought, and being unable to effect this they suspended judgment; and they found that quietude, as if by chance, followed upon their suspense, even as a shadow follows its substance. (Sextus Empiricus 1987, 12.29)

This mental quietude may well be the phenomenal signature of not-knowing and of uncertainty, coupled with a highly developed toler-

ance of ambiguity; and it may be intimately related to the ability to formulate a question or identify a problem while refraining from giving a solution.

What we can see now, especially by contrasting this point with what we said about intuitions earlier, is how mental quietude might be turned into a target for consciousness research in its own right, perhaps even forming a new branch of the psychology or cognitive neuroscience of interdisciplinarity. In particular, the mental state cultivated by the Pyrrhonian skeptics is diametrically opposed to that involved in intuitions. Both are phenomenal states only, and as such have no intrinsic epistemic warrant. However, where intuitions block further inquiry, mental quietude and the phenomenology of uncertainty promote it. The skeptic aims, in a sense, at a state in which inquiry has become permanent.

But there is also an important difference. Whereas intuitions and intuitive plausibility come to us naturally and effortlessly, open mindedness, the suspension of judgment and the tolerance of ambiguity are the result of careful cultivation, long-term practice, and sustained effort. From a purely evolutionary perspective, uncertainty and a non-judgmental attitude are costly and perhaps even dangerous, because they do not motivate action in the same immediate, quick, and unreflected way as intuitions.⁸

⁸ In fact, if doubt has an evolutionary function, it might be to prohibit activity and induce rest in situations in which the benefits of physical activity are outweighed by its risks, for instance in illness. Doubt and certainty of the theoretical sort may have more distinctly bodily precursors; they may be different ways of regulating how we relate to our own bodies and gauge our own level of physical ability. Carel (2013) describes bodily certainty as involving a tacit confidence “that our bodies will continue to function in a similar fashion to the way they have functioned in the past: we expect our stomachs to digest the lunch we have just eaten, our brains to continue to process information, our eyes to continue to see, and so on” (*ibid.* p. 4). By contrast, bodily doubt involves a breakdown of our beliefs about our own bodily capacities, but also a disruption on the level of subjective experience. “Bodily doubt is a physical sensation of doubt and hesitation arising in one's body. It is not solely cognitive, although it can be expressed in propositions. [...] Bodily doubt not only changes the content of experience, it also pierces the normal sense of bodily control, continuity, and transparency in a way that reveals their contingency. It shows our tacit faith in our own bodies to be a complex structure that becomes visible when it is disturbed. It changes the normal experience of continuity, transparency, and trust that characterize this structure” (*ibid.* p. 11). Bodily doubt is often associated with physical illness and depression, and in some cases, it seems this form of experiencing our own physical vulnerability may have a protective function. But according to Carel, the analysis of bodily doubt

If on encountering a bear in the wilderness you take too much time to contemplate the nature of the threat (or to question your intuitive assessment that the bear is indeed a threat), you might be eaten before you come to a conclusion. Clearly, introducing the Pyrrhonian spirit to such practical, everyday situations is absurd and perhaps even unhealthy. However, in the context of philosophical and scientific inquiry, cultivating vulnerability of the epistemic type (cf. Chinnery 2014) might be a strength and might help prepare the ground for genuine collaboration and fruitful discourse. But we can now also understand why, even in science, open mindedness is so frustratingly difficult to sustain: mental quietude is not a state of passivity or mental inertia. It is a mental ability that requires constant alertness and a lifetime of practice.

3.3 Acknowledging the problem of subjectivity

If open mindedness indeed draws from the same ideals as are rooted in Pyrrhonian skepticism, how can we put these insights to work in investigating phenomenal states and tackling the problem of subjectivity? In contemporary philosophy of mind, the problem of subjectivity is often taken to be the main conceptual and methodological obstacle for a true science of the mind. Can the first-person perspective be naturalized? What, exactly, is the place of subjectivity in the scientific world-view? And is there really something like “first-person data” that can—and perhaps must—enter the process of constructing a truly comprehensive theory of the conscious mind? Questions of this kind are

also illuminates the extent to which we are normally guided by a tacit and unshakeable kind of bodily certainty that typically cannot be rejected or rationally justified and that forms part of our brute animal nature. If this is right, then it might also explain why even the more abstract and theoretical variants of certainty continue to be associated with health and strength on the level of subjective experience—even though this confidence can be epistemically misleading. This also fits in nicely with the claim, elaborated in footnote 4, that ambiguity threatens the perceived coherence of the phenomenal self-model, whereas certainty, on the level of subjective experience, appears to be associated with heightened self-awareness. We might now say that doubt and the tolerance of ambiguity are an acquired taste: while in their early stages, they are often associated with discomfort or even anxiety, their cultivation may also be the key to genuine peace of mind.

good examples of high-level theoretical issues that require the epistemic virtues associated with an open mind. Even the editors of this collection have a tendency to disagree on this question—and we hope that this disagreement is of a constructive sort.

One of us (TM) thinks that a greater practical openness to so-called “first-person methods” on the part of researchers in philosophy and cognitive science alike might lead to great heuristic fecundity and would, perhaps dramatically, improve the quality and efficiency of research. Many such methods can be seen as the cultivation of a set of abilities that increase mental autonomy (M-autonomy; Metzinger 2013b, 2013d) and establish the inner preconditions for critical, rational thought: by stabilizing the first-person perspective, they create a more robust “epistemic agent model” (EAM; Metzinger 2013a, Box 1; Metzinger 2013d), or the experience of being a knowing self. At the same time he holds that there simply are no “first-person data” in any strict or conceptually more rigorous sense. Seriously assuming the existence of such data rests on an extended usage of a concept that is only well-defined in another (namely, scientific) context. First, the whole concept of a “first-person perspective” is just a visuo-grammatical metaphor, without a theory to back it up—and currently we simply don’t know what that could be, namely what “a” first-person perspective would look like (for a first conceptual differentiation, see Metzinger 2003, 2004; Blanke & Metzinger 2009). Second, “data” are extracted from the physical world by *technical* measuring devices, in a *public procedure* that is well-defined and well-understood, replicable, and improvable; and which is necessarily *intersubjective*. But in introspecting our own minds we never have any truly direct or immediate access to a mysterious class of “subjective facts”—all we have are neural correlates and publicly observable reports (which need not be verbal). Speaking of “first-person data” rests on an extended usage of a concept that is only well-defined in another context of application, rhetorically exploiting a fallacy of equivocation. “Data” are typically (though not always) gathered with the help of technical measuring

devices (and not individual brains) and by groups of people who mutually control and criticize each other's methods of data-gathering (namely, by large scientific communities). In particular, data are gathered in the context of rational theories aiming at ever better predictions, theories that—as opposed to phenomenological reports—are open to falsification.

To be sure, autophenomenological *reports*, theory-contaminated as they may be, are themselves highly valuable and can certainly be treated as data. But the experience “itself” cannot. However, even if one presupposes this rather straightforward view, having an open mind certainly also means acknowledging the additional fact that, for various reasons, this cannot be the *whole* story. It would be intellectually dishonest to deny without argument that what is sometimes called “first-person methods” could have enormous potential in our quest for a rigorous, empirically based theory of the human mind. The question rather is: What *exactly* is it about these methods that generates the extra epistemic value, if there really is one? It seems clear that not all epistemic virtues are *intellectual* virtues, and it is striking to note how such methods have played a central role in all cultures and in almost all ancient philosophical traditions of humankind. This is not only true for Asian systems of philosophy. At the very beginning of Western philosophy, Cicero (1971), in the *Tusculanae disputationes* (II 5), defined philosophy itself as *cultura animi*, as a way of caring for and cultivating the soul.

The other (JW) thinks that first-person data exist, and that for a true science of the mind, of consciousness and of subjectivity, it is important to acknowledge their existence. First-person data are not, however, to be found in the direct observation of conscious experience—thus far JW and TM are in perfect agreement—but in describing and more properly in reporting it. A first step towards seeing why this is the case is to clearly distinguish first-person reports from general opinions, convictions, or even intuitions about experience. First-person reports, in this view, are the product of (verbal or non-verbal) behaviors conducted with the sincere intent of conveying or recording certain rel-

evant information about a specific experience. They are not mere opinions about what it is typically like for oneself to have a certain kind of experience. They also should not be confused with attempts to generalize from one's own case to what it is typically like for other people to undergo a given type of experience, or with the practice, occasionally found in academic philosophy of mind, of relying on intuitive judgments or thought experiments to reach general conclusions about the necessary or even typical characteristics of given types of experience.

First-person reports, construed as sincere descriptions of specific and individual experiences, form the data-base of scientific consciousness research. They can be gathered with the help of public methods such as standardized interview techniques or questionnaires, and the data obtained from these reports are open to intersubjective validation (e.g., by using independent raters, different methods of statistical analysis and of scoring the content of reports, and so on). At the same time, this strategy works only against a background of trust that first-person reports can, when gathered under sufficiently ideal reporting conditions, be regarded as trustworthy with respect to the specific experiences they purport to describe. Indeed, assuming at least a subgroup of first-person reports to be trustworthy is a necessary condition of possibility for scientific consciousness research, for methodological reasons (see Windt 2013, 2015).⁹

Much of the serious work, in this view, will consist in identifying and improving the appropriate conditions under which maximally accurate experience reports can be obtained. Seen in this manner, the trustworthiness of first-per-

⁹ Clearly, this is not to say that such reports, or the data obtained from their analysis, are trustworthy with respect, for instance, to the neural underpinnings of the respective experiences, and we should not expect them to be. First-person reports, when gathered under ideal reporting conditions, are trustworthy with respect to the phenomenal character of experience only. Moreover, because this type of phenomenological information cannot be gleaned, for instance, from neuroimaging data, first-person data obtained from the analysis of experience reports necessarily complement third-person data. As dream researchers Tore Nielsen & Philippe Stenstrom (2005, p. 1289) put it, “[i]n an era of high-resolution brain imaging, similarly high-resolution reports of dream imagery may be needed”. A true science of consciousness will draw from different methodologies and different ways of measuring experience, and it will strive to integrate different types of data and different levels of description.

son reports becomes, to a considerable degree, a methodological problem for empirical research, not a principled philosophical or conceptual one, and the contribution of philosophy consists, at best, in showing why this is the case (again, see Windt 2013; for critical discussion, see Solomonova et al. 2014). By contrast, principled distrust in first-person reports, or even the attempt to investigate the phenomenology of experience independently of first-person reports, is an obstacle to a true science of consciousness.

While we, the editors, may disagree on the trustworthiness and epistemic status of first-person reports or even on the existence of first-person data in a strict sense, we certainly agree about the need to take our own subjective experience seriously, and we also agree that the epistemic stance we call “open mindedness” may well include a need to cultivate familiarity with our own subjective experience. In this respect, our accounts may well be complementary. Readers familiar with contemplative traditions may also have noted that there is a surprisingly direct and often quite literal correspondence between many classical notions such as “withholding judgment”, “mental quietude”, or “ataraxia”, and the practical instructions given by meditation teachers around the world, from different periods and different non-Western systems of philosophy. These notions are not only theoretical concepts—they draw our attention to the fact that there is more than one type of epistemic practice, and that open mindedness may in part be constituted by the set of abilities that connects them (Metzinger 2013c). On a more theoretical level, to have an open mind again means to acknowledge (and not repress) the fact that there may actually be a deep, unresolved ambiguity here, between the need to take subjective experience seriously and the suspension of judgment. In fact, bracketing one’s own folk-psychological or intuitive judgments about experience is part of what it takes to move towards a truly scientific approach to subjective experience. For this reason, open mindedness involves cultivating not only a particular attitude towards one’s beliefs, but also towards oneself as a believer.

A similar tolerance of ambiguity is at play in the attitude of lending equal credence to reports from different subjects, acknowledging inter- and intrasubjective variation in experience, and, ultimately, trying to integrate these reports into a maximally large data-base, while resisting the pull of generalizing from one’s own case or engaging in armchair phenomenology (where this involves pumping intuitions about experience rather than carefully observing and describing what it is like to have particular experiences). We might even say that this strategy of stepping back from one’s own convictions about experience and formulating questions about the phenomenal character or the subjectivity of experience is in keeping with the Pyrrhonian spirit: both are directed at academic disputes and assume commonplace experience or individual experience reports to be trustworthy, and both strive towards a confrontation of theoretical statements with real-world counterexamples, with the aim of ultimately giving rise to more sophisticated theories.

The issue of subjectivity is an excellent example of a persevering problem that comes in many different guises and reappears on many different levels. Perhaps there really *is* something about the conscious mind that cannot be explained reductively, even in principle. But searching for a maximally parsimonious scientific explanation is a rational research heuristic, not an ideology. It should never be a substitute for religion, and as such it carries it with it no immediate metaphysical commitments. To have an open mind is an *epistemic* stance, which means that epistemic progress is what counts in the end. Many of the authors in this collection, including the editors, are staunch methodological naturalists, because they view philosophy and science as engaged in essentially the same enterprise, pursuing similar ends and using similar methods. If it could be shown, however, more precisely than ever before in the history of philosophy and science, that there are strictly irreducible aspects of the human mind, then most of the authors in this collection, and indeed most researchers in this field, would be satisfied with this result. They would have what they wanted all along: epistemic progress.

4 The wider context

Having an open mind means never losing sight of the bigger picture and being continuously aware that scientific research, including research on the mind, is embedded in a wider context. In what follows, we will very briefly draw attention to three examples of what we mean by the “bigger picture” and the “wider context”: ethical, anthropological, and sociocultural issues; globalization and transcultural philosophy; and what we provisionally call “the sapiential dimension”—getting *philosophy* back into philosophy.¹⁰ Let us begin with the ethical ramifications of the type of work presented in this collection.

New theories lead to new technologies and new potentials for action. Gradually, they also change the image of humankind, a fact that may in turn have major social and cultural consequences. Having an open mind means being sensitive to normative issues and ethical aspects of research in philosophy of mind and cognitive science. It also means acknowledging the fact that the human mind is a culturally embedded phenomenon and that what we come to believe about it will eventually change not only sociocultural practice, but our own minds as well. Such “soft issues” are not empirically tractable, at least not in any direct manner (Metzinger 2000, pp. 6–10; Metzinger 2009). Here, perhaps even more so than elsewhere, the challenge is to formulate the right kinds of questions in a rigorous, precise, and fully intelligible manner. These questions are certainly difficult, but they are also clearly *relevant*.

4.1 Sensitivity to ethical issues

Theoretical innovation leads to technological innovation, necessitating careful and reflected risk

¹⁰ Again, this comes back to the classical idea of wisdom as not only knowing how to live well, but also succeeding at doing so (Ryan 2014). There is also a clear connection between open mindedness as an epistemic practice and its ethical dimension. As Russell (1912/1999, p. 116) puts it, “[t]he mind which has become accustomed to the freedom and impartiality of philosophic contemplation will preserve something of the same freedom and impartiality in the world of action and emotion. [...] The impartiality which, in contemplation, is the unalloyed desire for truth, is the very same quality of mind which, in action, is justice, and in emotion is that universal love which can be given to all, and not only to those who are judged useful or admirable”. The true value of philosophy lies not just in its effects on our thoughts, but on our lives, on our actions; “it makes us citizens of the universe” (*ibid.*, p. 116).

assessment. For example, modern virtual reality technology not only enables the concrete realization of a large number of new experimental paradigms, but has also provided us with many novel and philosophically relevant insights into the multimodal bodily foundations of selfhood and subjectivity (Blanke 2012; Blanke & Metzinger 2009; Metzinger 2014). In combination with constantly improving brain-computer interfaces, virtual reality technology also possesses the potential for military applications, for example via *virtual or robotic re-embodiment*. New ways of causally coupling the human-self-model with avatars and surrogate bodies in virtual reality will have clinical benefits in the medical treatment of patients and, perhaps, in rehabilitation programs for prisoners. But it also opens the door to new forms of consumer manipulation and potentially unexpected psychological side-effects (e.g., Blascovich & Bailenson 2011).

A second example of the social and political dimension of new action potentials, in terms of how they might intervene in the brain, is provided by new developments in pharmaceutical cognitive enhancement (Merkel et al. 2007; Metzinger & Hildt 2011). Cognitive enhancement is a molecular-level technology, which aims to optimize a specific class of information-processing functions: *cognitive* functions, physically realized by the human brain. The human brain, however, is also embodied as well as embedded in a dense network of environmental interactions, many of which are of a distinctly cultural and social nature. And it not only possesses a long evolutionary history, but also changes over an individual’s lifespan. Here, the central philosophical problem is that *normative* elements are already built into the concept itself. In bioethics, the term “enhancement” is “usually used [...] to characterize interventions designed to improve human form or functioning beyond what is necessary to sustain or restore good health” (Juengst 1998, p. 29). As opposed to medical treatments or therapies, enhancements modify physical or mental characteristics in healthy individuals, just like cosmetic surgery. In psychopharmacological enhancement, psychoactive drugs originally devised as therapy

for specified diseases are typically used off-label or illicitly by normal, healthy individuals in order to modify brain functioning. In the future, how exactly can we benefit from scientific progress, for example by influencing and constructively interacting with the ever-developing neuronal architecture of our brains on a molecular level, while not leaving the social context out of consideration?

Who counts as a “healthy individual”? A trivial but important point is that concepts like “normal mental functioning” or, say, “normal age-related cognitive decline” possess a statistical and a normative reading. The semantics of both types of concepts change over time. For example, the statistical and descriptive features of “normal mental functioning” or “normal age-related cognitive decline” change as science progresses, as the predictive success of our theories improves, and as textbook definitions are adapted. Our concepts become richer in content and more differentiated. But if a specific society suddenly has new tools and new potentials for action—say, to alter certain cognitive functions in the elderly—then the statistical distribution of even those objective properties underlying a purely statistical notion of what is “normal” may also change. Cognitive enhancement is a neurotechnology, and technologies change the objective world. However, objective changes are also subjectively perceived and may lead to correlated shifts in value judgments. Concepts such as “healthy individual”, “normal mental functioning”, or “normal age-related cognitive decline” always have a descriptive as well as a normative reading, because they appear in statements about what human beings *should* be like. Is it really necessary to succumb to memory loss or a decreasing attention span after the age of 55? If other options are actually on the table, does this turn passively capitulating to age-related cognitive decline or certain individual limitations in the ability to engage in high-level, abstract thought into a cognitive form of unkemptness and dishevelment?

In this example, the not-so-trivial challenge lies in understanding the dynamic interaction between “normality” (in the descriptive sense) and “normalization” (in the normative

sense). The theoretical and social dynamics linking both concepts and their interpretation is highly complex. It involves scientific theories (in cognitive neuroscience, molecular neurobiology, and psychopharmacology), applied philosophical ethics, changing cultural contexts, globalization, policy-making, as well as industrial lobbies trying to influence the historical change of our very own concepts and their meaning in order to market new products. Normalization is a complex sociocultural process by which certain new norms become accepted in societal practice. For this reason, the scientific process, say, of optimizing textbook definitions, empirical predictions, and therapeutical success has a political dimension as well. It attempts to firmly ground theoretical entities such as “normal mental functioning” or “normal age-related cognitive decline” in empirical data, but it is also driven by individual career interests, influenced by funding agencies, the pharmaceutical industry, media coverage, and so on.

A third important example of how new ethical issues emerge is presented by the question of animal consciousness and animal suffering. What is the ethics of creating suffering in non-human species, for example in the scientific pursuit of uniquely human epistemic goals? Much recent research shows that many animals are very likely not only conscious, but also self-conscious and able to suffer (Brown 2015; Boly et al. 2013; Edelman & Seth 2009; Seth et al. 2005). They represent a frustration of their own individual preferences on the level of their consciously experienced self-model and thus *own* their sensory pain. They are also very likely to be unable to distance themselves from negative emotions such as fear, anxiety, or depression. In the light of new and better descriptive theories of consciousness, classical normative issues such as animal ethics reappear in a new guise and with increasing urgency.

Philosophical questions such as “Who or what exactly should count as an object of ethical consideration?” soon may also become relevant for the applied ethics of synthetic phenomenology, that is, for all research programs in artificial intelligence that risk or even directly intend the creation of phenomenal experience,

of truly subjective, conscious states in non-biological hardware. “Synthetic phenomenology” (SP) was first introduced by J. Scott Jordan in 1998, explicitly paralleling the idea of “synthetic biology”.¹¹ The possibility of machine consciousness now is not only part of the bigger picture and the wider context mentioned above, it also illustrates how theoretical innovation may eventually lead to technological innovation and require a careful assessment of possible risks. For example, the *Principle of Negative Synthetic Phenomenology* (Metzinger 2013b, pp. 2–8) is an ethical norm that demands that, in artificial systems, we should not risk the unexpected emergence of conscious states belonging to the phenomenological category of “suffering” or even aim at the direct creation of states that would increase the overall amount of suffering in the universe. But how exactly are we to unpack the logical details of this normative proposal? How does one approach these new types of questions in a rational and data-driven manner? Machine consciousness, just like VR-technology, pharmaceutical enhancement, and animal suffering is another example of a topic where a lack of imagination might prove dangerous and where an open-minded approach is pertinent.

Perhaps one central aspect of this problem is that in an increasing number of cases we will not only have to ask, “What is a good action?” but also, “What is a good state of consciousness?” Opening, cultivating and further developing one’s own mind clearly is in the spirit of not only Cicero, Plato, and the ancients—systematically increasing our own mental autonomy seems to be a common ideal shared by many of humankind’s philosophical traditions. However, the boundary conditions for this old philosophical project are beginning to change because the tools for manipulating or

even systematically cultivating our own minds are constantly becoming better—and precisely as a result of interdisciplinary, empirical work in the Mind Sciences. If we arrive at a comprehensive theory of consciousness, and if we develop ever more sophisticated tools to alter the contents of subjective experience, we will have to think hard about what a *good* state of consciousness is. This again illustrates the point that as some parts of neurotechnology inevitably lead to consciousness technology, new normative issues arise and classical philosophical questions reappear in new guises (Metzinger 2009).

As editors of this collection, we do not want to take a specific position on any of these important and highly controversial issues. We merely want to point out that having an open mind also means cultivating a specific kind of sensitivity: a sensitivity for the actual and potential suffering of other sentient beings, for newly emerging ethical issues and for the obvious fact that the kind of research we are developing together does not take place in a political, social, or cultural vacuum. For example, open mindedness also requires a self-critical sense of responsibility to global society as a whole. It is also in *this* context that new conceptual bridges have to be built between artificial intelligence, cognitive neuroscience, philosophy of mind, and ethics. Once more, a first and important step may be to carefully consider the questions themselves, rather than to rush into an answer or attempt to quickly implement mere technocratic solutions. Ultimately, all of these questions have a lot to do with the classical philosophical problem of what a good life actually is.

4.2 Globalization and intercultural philosophy

There is not only an ethics of science, there is also an ethics of globalization. It has to do with fairness and, for example, the willingness of the rich to relinquish some of their sovereignty for the benefits of cooperation. Of course, there are technical issues behind philosophical notions such as “global fairness”. But many would agree that we should distribute resources in a way that helps the worst-off, and that the only way

¹¹ See Chrisley 2009, p. 68 and Chrisley & Parthemore 2007, note 2. SP encompasses a variety of different approaches, methodologies, and disciplines, but what they all have in common is that they see SP as the construction or guided dynamical self-organization of phenomenal states in artificial systems. They also share the deep-seated methodological intuition that any scientific explanation of consciousness necessarily involves a systematic *re-construction* of the target phenomenon. See Gamez (2008, pp. 887–910); Holland & Goodman (2003); Holland et al. (2007); Chrisley & Parthemore (2007); Aleksander (2008) for a first overview.

of justifying giving more to those members of humanity who are already well-off is if it demonstrably improves the position of those in the poorest and most dangerous parts of the world as well. The movement of effective altruism uses scientific research to determine the optimal ways of distributing goods to the poorest regions of the world, with the goal of maximizing the benefits and long-term efficiency for instance of donations to charities (for general information, see <http://www.effectivealtruism.org/>). Such debates apply to the globalization of science and philosophy as well. In this context, it is interesting (and sobering) to note how in academic philosophy, the basic idea of making scholarly work available free of charge and free of usage restrictions online is vastly underdeveloped in comparison to other fields of research. It is also sobering to note that academic philosophy, possibly more than other academic disciplines, continues to be dominated by white, Western (and mostly Anglo-Saxon) males. This is not just reflected in philosophy departments themselves, but also in well-known and widely consulted ranking systems, which almost exclusively focus on Anglo-Saxon departments. We could do much better here, in all of these respects. Of course, many of us have long realized this, and as editors of this collection, we are preaching to the choir. What is needed now are viable ways of changing this situation.

Because of the open access format of the Open MIND collection, which was conceived of, in part, as a donation of intellectual property, we want to focus on one single aspect here. One might argue that the current subscription-based publishing system, which comprises nearly all of the top-ranked journals that young researchers in particular strive to have on their CVs, is inherently conservative, stabilizes the academic status quo, and, given the context of academic globalization plus the urgent need to strengthen deeper and not just intellectual forms of intercultural exchange, potentially leads to a “global closed-mindedness”, to a narrowing of intellectual and scholarly life. Typically, publically funded academics will be involved on different levels and in different stages of the publication process, not only as authors, but also as review-

ers, members on editorial boards, editors, and so on. Indeed, these types of participation are awarded and often expected by hiring committees. Yet, despite all of the hours of free labor (from the perspective of the publishing houses), the scientific publications that flow out of this process are often locked behind a paywall, giving authors only limited rights to distribute their own research. More innovative journals give authors the opportunity to publish their papers open access—typically in return for a hefty publication fee that, once more, is most likely to be funded by rich universities in affluent countries. Again, we can, and should, do much better.

Through their work, scientists and philosophers continuously produce knowledge and new intellectual property. However, there exists not only knowledge production, but also knowledge consumption—and the overall process has an economical basis. How should such goods be justly distributed? Who can *participate* in the process of producing and consuming them? The world continues to be divided into “haves” and “have-nots” when it comes to accessing the fruits of the intellectual labor of humankind. The point is not only that taxpayers should have access to the results of all publicly funded work. A more central point is that, given globalization, we now need a much more transcultural type of philosophy. In order to realize this goal, we urgently need to experiment with different formats of open access publishing, testing out what works best. In this way, we could finally create a unified public sphere for research—a “global workspace” for the science and philosophy of all humankind. Clearly, this in itself is not sufficient, but is a very first, necessary step.¹² Still, the historical transition we are witnessing is one where having an open mind also

¹² And new questions continuously arise. Is it, for instance, unethical to publish one’s research in scientific journals or books that are not open access and which therefore systematically exclude a large majority of students and researchers from the less affluent part of the world? If you answer affirmatively to this question, would you also say that it is unethical to consume research published in books or journals that are not open access? And do you think, in terms of civil disobedience, that it is permissible to disregard copyrights (and authors’ rights to royalties) to make such research, either your own or even that of others, openly available? This is just a small selection of the potentially difficult questions facing today’s scholars and researchers. And people are already acting upon their answers (see, for instance, Ludlow 2013).

means publishing open access whenever possible—which in no way excludes *additionally* using, and paying for, traditional dissemination formats as well. But in creating humanity's global workspace, as Steven Harnad (2007) puts it, it has now simply become “unethical for the publishing tail to be allowed to continue to wag the research dog.” What is needed is an honest and objective assessment of the most effective methods of scientific publishing—where effective not only means cost-efficient from the perspective of large publishing houses, but also addresses the dual challenges of optimizing the quality of research and peer-review processes while making scientific results available to all interested researchers and scholars.

“Intercultural philosophy” may sound good—but what does it really mean? Philosophy was born at different places and at different times, for example in India, in China, and in Europe. Philosophical thinking evolved in different cultural contexts that were often quite independent of each other and sometimes remained largely isolated for many centuries. Globalization now forces us to face the need to create novel forms of communication between philosophers as well as new forms of cooperation between different traditions and cultures. Yet this development is also an opportunity. The idea of “intercultural philosophy” is certainly not new, and there are many different ways of spelling it out. Here, we want only to point out that in our view, intercultural philosophy should not be a new academic discipline, but that it is, again, an *attitude*, an increasingly important form of epistemic practice.

At the same time, not *all* philosophical research contexts originally evolved in isolation, and the globalization of wisdom may be older than we think. To give just one familiar example, it is noteworthy that Pyrrhonian skepticism plausibly has a strong (and entirely mutual) intercultural dimension as well. The practice of using standardized arguments involving opposing statements to cultivate positionlessness, suspension of judgment, and epoché can be found in the Indian tradition as well, for instance in the Madhyamaka tradition and in Nagarjuna's writings. Textual evidence suggests that not only might Pyrrho himself have

been inspired by ideas with which he came into contact in India, but also that later, Sextus's version of Pyrrhonian skepticism might have shaped Nagarjuna's Middle Way (Dreyfus & Garfield 2010; Geldsetzer 2010; Kuzminski 2008). Having an open mind, in this sense, involves not only bridging *disciplinary* cultures, but also integrating different research traditions from different cultures and different periods and looking for their common sources.

Obviously, the open-minded “pooling of intellectual resources” that we mentioned above must increasingly also include philosophers not only from Europe or the Anglo-Saxon world. From a traditional Western perspective, epistemic humility also means acknowledging that other philosophical traditions may long ago have had deep insights into theoretical problems that still puzzle us today, even if their knowledge is not presented in a format and terminology that we are used to or can easily understand. It would be intellectually dishonest to assume that the style of thought developed in Anglo-Saxon analytical philosophy is the only way of being intellectually honest. And obviously, if, as we do, one calls for an expansion of the principle of charity into interdisciplinary discourse, then one should also accept that the same principle applies to intercultural collaboration. If there is to be a culture of charity, then it must be a *global* culture of charity—including open access publishing and global fairness in the distribution of academic goods. Today, even more than in the past, this is another reading of what it means to have an open mind.

4.3 The sapiential dimension

Thanks to the internet and major technological advances, modern academic life is unfolding at a greater pace than ever before. It has also become more competitive than it ever was in the past. This development bears the promise of progress; but it also poses a very real risk. As knowledge production becomes a commodity and academia is increasingly reorganized based on economic principles of marketing and business administration, universities are replacing tenure-track lines with adjunct teachers and a constantly growing number of brilliant young academics are now

competing for scarce resources in a globalized academic environment. The acceleration of academic life as well as increased social pressure are beginning to have psychological effects on individual researchers as well. A recent surge in the detection of fraud and scientific misconduct may be a sign of underlying counterproductive incentives that have begun to influence scientists worldwide. According to a report in the journal *Nature*, published retractions in scientific journals have increased by around 1,200% over the past decade, even though the number of published papers grew only 44% in the same period (Van Noorden 2011). A detailed review of all 2,047 biomedical and life-science research articles indexed by PubMed as retracted by the 3rd of May, 2012 revealed that only 21.3% of retractions were attributable to error (Fang et al. 2012). 67.4% of retractions were attributable to misconduct, including fraud or suspected fraud (43.4%), duplicate publication (14.2%), and plagiarism (9.8%). It is also possible, however, that the rising number of retractions has been caused by a growing propensity to retract flawed and fraudulent papers and does not in fact involve a substantial increase in the prevalence of misconduct (Fanelli 2013). These numbers might therefore also suggest an increasing willingness to retract faulty publications. They might also be artefacts of an increased availability of data on such retractions. We do not know what the final interpretation of such data should be. But we do regard them as one potential indicator of overheated competition turned counterproductive.

In philosophy, there is a high and continuously growing pressure for specialization, and this historical development presents a major problem. One classical model of what philosophy is says that philosophers are “specialists for the general”, who are concerned with integrating the knowledge of their time into an overarching conceptual model. As one German idealist philosopher put it, philosophy “is its own time comprehended in thought”.¹³ Today, the realization of this metaphilosophical vision has long become an impossible task for even the

greatest scholar. The sheer number of publications in any given, specialized area of research—such as embodied cognition, self-consciousness, or the evolution of culture and complex societies—has become so large that it is now extremely difficult for any ambitious young philosopher to even get an overview of the field. At the research frontier, great progress has been made in the fine-grained differentiation of research questions, while conceptual precision, argumentational density, and the general speed with which technical debates are conducted is continuously rising. This historical shift has become particularly obvious in philosophy of mind. In the age of cognitive neuroscience and Bayesian modeling, “raising one’s own age to the level of thought”, as Hegel put it, has simply become an impossible task. On the other hand, philosophers of mind are not embedded journalists of the neuroscience industry. A philosopher’s task today clearly goes far beyond offering methodological criticism plus a bit of applied ethics. Philosophers should not confine themselves to laying and clarifying some conceptual foundations or just developing a local, domain-specific “conceptual commentary” on the general way in which the empirical Mind Sciences change our perspective on reality and the human mind’s position within it. In the future, philosophers must more actively introduce their own epistemic goals into the overall process as well. Failure to do so is to exercise a counterproductive sort of epistemic humility—and runs the risk of letting academic philosophy slip into irrelevance.

Having an open mind also means that there are no taboo topics. At the outset, philosophy was the “love of wisdom” and, as everybody knows, knowledge and wisdom are not the same thing. Knowledge is something that can be accumulated in an incremental and systematic way, but wisdom has to do with synthesizing very different kinds of knowledge in ways that are practically relevant, for example with respect to knowing what a good life is and, importantly, also with being *successful* at living a good life (Ryan 2014). This in turn may include actively minimizing the number of unjustified beliefs one has and continuously maximizing the

¹³ Hegel, in his preface to the *Elements of the Philosophy of Right*, ed. Allen W. Wood, trans. H. B. Nisbet, Cambridge, UK: Cambridge University Press, 1991.

dynamic coherence between one's beliefs, one's values, and one's actions. Perhaps wisdom can also be characterized by a sustained striving for accuracy and for the possession of a wide variety of epistemically justified beliefs on a wide variety of relevant subjects—with one such subject being the deep structure of the human mind itself. In this case, knowledge will automatically be self-knowledge, and the question now becomes on what level the *relevant* form of self-knowledge is to be found. Tackling this problem may involve a commitment to a deeper form of rationality that includes not only epistemic humility, but heightened sensitivity towards moral issues and one's limitations in both fields.

It now has become dramatically obvious that something has been lost along the way. Academic life has become distinctly *unphilosophical*. Professionalization, acceleration, and excessive competition have led us into a form of academic life that can now very rarely be described as a good life. First, it seems safe to say that many of the best and leading researchers are not very successful at living a good life—even if they are philosophers who, at least at the beginning of their careers, may have had a great interest in exactly what a “good” life in the philosophical sense might be. Second, overheated competition increasingly draws people into the field who are predominantly interested in competition and professional success *per se*, and not so much in the pursuit of knowledge, let alone wisdom. But intellectual superiority and insight are different things, just as knowledge and wisdom are. There is no intrinsic link between striving for intellectual superiority and being intellectually honest, practicing epistemic humility and cultivating an atmosphere of charitable collaboration. In academic philosophy, the sapiential dimension, in which theoretical insight and practical know-how are deeply interwoven, has now been lost almost completely, and one aspect of what it means to have an open mind—as opposed to just being professional, knowledgeable, and smart—is to be aware of this fact and to be ready to face it.

We, the editors, certainly do not claim to know what exactly philosophy really is or what

it is to lead a good life, nor do we always agree on these questions—but we are convinced that whatever the answer is, it is deeply connected with a particular kind of attitude that reaches back all the way to the skeptical tradition, East and West. Philosophy at its best is not just purely academic or technical: it is also a practice, a way of life; and its theoretical and practical dimensions should never be completely independent of each other. This is what we mean when we say that academic philosophy would greatly profit from a sapiential dimension. And if we are right to say that philosophy is, among other things, an epistemic practice, a particular style of thinking resulting from the cultivation of an open-minded attitude (and one that is skeptical, we might add, in the most constructive sense), then this may also suggest a new reading of what it means to say that philosophy has an important role to play in the Mind Sciences. Asking for an interaction between cognitive neuroscience and philosophy as academic disciplines is one thing—but asking for the introduction of a particular way of thinking and a particular type of collaborative practice—a more genuinely *philosophical* attitude—into scientific research is another. We hope that by now it is clear that we think philosophy can contribute to the Mind Sciences in both respects, as an academic discipline and as an epistemic practice. Still, what we have been discussing here under the heading of open mindedness is first and foremost an example of philosophy as an epistemic practice—and as such it can be quite independent of philosophy as an academic discipline. Indeed, this is why we think that an important goal is to put philosophy, in this practical and classical sense, back into philosophy in the academic sense as well.

We openly admit that we have no ready-made answer to the question of how to re-introduce the sapiential dimension into modern academic philosophy, in a way that is rational and intellectually honest. In fact, we think this might well be the biggest challenge for the future. Obviously, what we call the “sapiential dimension” here has nothing to do with any kind of theology or organized religion. And we suspect that the real value of what we called “first-

person methods” above may lie not in supporting dubious metaphysical arguments, but lies, in part, in their potential for reintroducing the sapiential dimension into academic philosophy. But we also want to point out that this could simply be empirically false. Sometimes it is enough to remain with the question, to simply see it for what it is and to face the facts. Sometimes things take care of themselves. As we said when sketching the problem of subjectivity, to have an open mind means to acknowledge (and not repress) the fact that there may actually be a set of deeper metatheoretical ambiguities here. Having an open mind can also consist in admitting the existence of a problem—and that is all we want to do here.

4.4 Developing new forms of interdisciplinarity

Taking empirical constraints into account has become absolutely central in current philosophy of mind. However, there are different models of what good interdisciplinary practice is and *how* empirical constraints are to be satisfied or integrated. Interdisciplinary philosophy of mind does not simply consist in turning away from old-school armchair philosophy, which sometimes took intuitions as main input for philosophical work. And it would be false to say that “pure” philosophy has no place in the newly unfolding scheme of things—there is clearly relevant and highly valuable work that has only a small empirical component, or perhaps even none at all. One aspect of the Open MIND approach is that young philosophers should increasingly become active as experimenters themselves, for instance by proposing epistemic goals and novel experimental designs to empirical researchers and even by joining their colleagues from different disciplines to work on shared research projects. Another aspect of the approach, as we noted earlier, is that the extended principle of charity applies not only to the relationship between disciplines, but also to that between different generations of researchers.

We are all learning as we go along. Perhaps most centrally and most obviously, to have an open mind means to acknowledge the fact that while there has long been an “interdiscip-

linarity turn” in philosophy of mind, the real task consists in creatively testing out and developing entirely new *types* of interdisciplinary cooperation. For example, it is important to preserve a critical spirit and an openly inquisitive mindset—interdisciplinarity must never be purely decorative, a fashionable necessity, or reduced to a rhetorical element in edifying Sunday speeches. Along the way, we will also need a new understanding of progress, of acceptable forms of inquiry and methods, as well as new measures of success, for instance concerning novel forms of collaboration and publication formats that are still under the radar of institutionalized impact factors.

To give a second example, the newly emerged discipline of neuroethics is an important and innovative form of interdisciplinary philosophy, but it should never indirectly contribute to moral hypocrisy, as a fig leaf ultimately used by others to cover the failure to directly and open-mindedly address the political issues involved. If interdisciplinarity becomes merely strategic (e.g., in dealing with funding agencies) or is really guided by off-topic motives, then it loses its systematic force and becomes counterproductive and stale. Interdisciplinary philosophy of mind is not simply about being empirically informed, or about introducing strong and fine-grained “bottom-up constraints” in the formation of new theories about mind and consciousness. It may actually be about the emergence of a new type of researcher. We like the idea of “dyed-in-the-wool interdisciplinarity”, where “dyed-in-the-wool” is not used in a pejorative sense but indicates that young philosophers have learned how to *think* in a way that transgresses boundaries between disciplines, naturally and effortlessly. The classical approaches were intuition-based, and they made analytical philosophy one of the strongest intellectual currents of the 20th century. But we are now slowly moving from a priori methods and thought experiments to real experiments, and from abstract metaphysical questions about the relationship between mind and body to the investigation of specific aspects of cognition (Knobe 2015). And while it is clear that an open-minded philosophy of mind should not be

strictly or exclusively data-driven, it is equally true that it should be both empirically informed and informative, guided (but not completely constrained) by empirical data and theoretical-conceptual considerations alike.

In the end, there is also a sociological aspect to the current transition in our understanding of what good philosophy amounts to. [Max Planck](#), the German theoretical physicist who created quantum theory and won the Nobel Prize for Physics in 1918, famously said: “A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it” (1948). As the editors of a collection promoting, among other things, senior–junior interaction, we think this may be a bit too pessimistic—and once more, we leave it to our readers to decide how successful this interaction was here, in this project. Still, for now, a careful suggestion is that possibly, the old should learn a little more from the young.

One of our experiences with the MIND Group was that there was a difference between what one might call “junior mentoring” and “senior mentoring”. Junior researchers need friends in neighboring disciplines whom they can trust and ask about literature, current trends, and technical issues that are hard to understand. Our experience is that interdisciplinary exchange works best in excellent young people who are not yet on the job market, and in non-competitive situations in which at best no holders of academic resources are present, such as senior researchers who have grants, post-doc positions, etc. to give away. Good and established systems of senior–junior mentoring already exist, but we believe that given the current situation, junior–junior mentoring is an important resource to be developed as well. For this reason, in the Open MIND project, we installed a form of junior–junior mentoring during the anonymous peer-review process for commentaries. And while replies can be seen as a form of senior–junior mentoring, there was also, covertly in the form of target article reviews, a phase of junior–senior mentoring, in which some of our junior members not only wrote their first

reviews ever, but now, after the collection’s publication, can also see for themselves how their comments were implemented and whether this maybe even led to an improvement of the target papers. But above all, it is important that young people from the *same* generation have the opportunity to meet each other and form their own, autonomous networks based on shared interests and mutually shared (or acquired) expertise. And this will require a radical restructuring of research funding and of the university system itself, as well as new subsidizing schemes. The function of older, more mature researchers may rather consist in creating and offering such platforms, giving a better overview of the intellectual landscape and offering insight into what is really relevant in a specific phase of a young researcher’s academic life. Today, the sociological aspect of what it means to have an open mind has an unprecedented global dimension. In trying to promote young blood, mostly in Germany, we found that language and cultural barriers actually are often higher than we wanted to admit. If what we have said about the ethics of globalization and intercultural philosophy here is correct, then we might not only need new formats of interdisciplinary and intragenerational collaboration, but also new types of intercultural mentoring as well.

As we said at the outset, instead of an introduction we wanted to begin a new conversation by offering some first starting points and perhaps even first building blocks for a fresh understanding of what, today, it could mean to have an open mind. Once again, we openly admit that we have no ready-made answers. But we are convinced that it is important to ask these questions. Somehow, we have to get philosophy back into philosophy.

Acknowledgements

We would like to thank Michael Madary, Nicole Osborne, Marius Jung, and Daniela Hill for editorial support and their helpful comments on an earlier version of this manuscript. And, as always, we are deeply indebted to Stefan Pitz, Anja Krug-Metzinger, and Janice Kaye Windt for their support.

References

- Aleksander, I. (2008). Machine consciousness. *Scholarpedia*, 3 (2), 4162-4162. [10.4249/scholarpedia.4162](https://doi.org/10.4249/scholarpedia.4162)
- Alexander, J. (2012). *Experimental philosophy: An introduction*. Cambridge, UK: Polity.
- Alexander, J. & Weinberg, J. M. (2007). Analytic epistemology and experimental philosophy. *Philosophy Compass*, 2 (1), 56-80. [10.1111/j.1747-9991.2006.00048.x](https://doi.org/10.1111/j.1747-9991.2006.00048.x)
- Bealer, G. (1998). Intuition and the autonomy of philosophy. In M. R. DePaul & W. Ramsey (Eds.) *Rethinking intuition. The psychology of intuition and its role in philosophical inquiry* (pp. 201-239). Boston, MA: Rowman & Littlefield.
- Bett, R. (2014). Pyrrho. *The Stanford Encyclopedia of Philosophy*. <http://plato.stanford.edu/archives/win2014/entries/pyrrho/>
- Blanke, O. (2012). Multisensory brain mechanisms of bodily self-consciousness. *Nature Reviews Neuroscience*, 13 (8), 556-571. [10.1038/nrn3292](https://doi.org/10.1038/nrn3292)
- Blanke, O. & Metzinger, T. (2009). Full-body illusions and minimal phenomenal selfhood. *Trends in Cognitive Sciences*, 13 (1), 7-13. [10.1016/j.tics.2008.10.003](https://doi.org/10.1016/j.tics.2008.10.003)
- Blascovich, J. & Bailenson, J. N. (2011). *Infinite reality - Avatars, eternal life, new worlds, and the dawn of the virtual revolution*. New York, NY: William Morrow.
- Boly, M., Seth, A. K., Wilke, M., Ingmundson, P., Baars, B., Laureys, S., Edelman, D. B. & Tsuchiya, N. (2013). Consciousness in humans and non-human animals: Recent advances and future directions. *Frontiers in Psychology*, 4 (625), 1-20. [10.3389/fpsyg.2013.00625](https://doi.org/10.3389/fpsyg.2013.00625)
- Brown, C. (2015). Fish intelligence, sentience and ethics. *Animal Cognition*, 18 (1), 1-17. [10.1007/s10071-014-0761-0](https://doi.org/10.1007/s10071-014-0761-0)
- Cappelen, H. (2012). *Philosophy without intuitions*. Oxford, UK: Oxford University Press.
- Carel, H. (2013). Bodily doubt. *Journal of Consciousness Studies*, 20 (7-8), 7-8.
- Chignell, A. (2010). The ethics of belief. *The Stanford Encyclopedia of Philosophy*. <http://plato.stanford.edu/archives/fall2010/entries/ethics-belief>
- Chinnery, A. (2014). On epistemic vulnerability and open-mindedness. *Philosophy of Education Archive*, 63-66.
- Chrisley, R. (2009). Synthetic phenomenology. *International Journal of Machine Consciousness*, 1 (1), 53-70. [10.1142/S1793843009000074](https://doi.org/10.1142/S1793843009000074)
- Chrisley, R. & Parthemore, J. (2007). Synthetic phenomenology: Exploiting embodiment to specify the non-conceptual content of visual experience. *Journal of Consciousness Studies*, 14 (7), 44-58.
- Chudnoff, E. (2013). Intuitive knowledge. *Philosophical Studies*, 162 (2), 359-378. [10.1007/s11098-011-9770-x](https://doi.org/10.1007/s11098-011-9770-x)
- Cicero, (1971). *Tusculan disputations*. Cambridge, MA: Harvard University Press.
- Clifford, W. K. (1999). The ethics of belief. In T. Madigan (Ed.) *The ethics of belief and other essays* (pp. 70-96). Amherst, MA: Prometheus.
- Crane, T. (2014). The problem of perception. *The Stanford Encyclopedia of Philosophy*. <http://plato.stanford.edu/archives/win2014/entries/perception-problem/>
- Dennett, D. C. (1991). *Consciousness explained*. New York, NY: Little, Brown and Company.
- (2013). *Intuition pumps and other tools for thinking*. New York, NY: W. W. Norton & Company.
- Diogenes Laertius, (1943). *Lives of eminent philosophers*. Cambridge, MA: Harvard University Press.
- Dreyfus, G. & Garfield, J. L. (2010). Madhyamaka and classical Greek skepticism. In G. Dreyfus, B. Finnigan, J. L. Garfield, G. M. Newland, G. Priest, M. Siderits, K. Tanaka, S. Thakchoe, T. Tillemans & J. Westerhoff (Eds.) *Moonshadows: Conventional truth in Buddhist philosophy* (pp. 115-130). New York, NY: Oxford University Press.
- Edelman, D. B. & Seth, A. K. (2009). Animal consciousness: A synthetic approach. *Trends in Neurosciences*, 32 (9), 476-484. [10.1016/j.tins.2009.05.008](https://doi.org/10.1016/j.tins.2009.05.008)
- Fanelli, D. (2013). Why growing retractions are (mostly) a good sign. *PLoS Medicine*, 10 (12), e1001563. [10.1371/journal.pmed.1001563](https://doi.org/10.1371/journal.pmed.1001563)
- Fang, F. C., Steen, R. G. & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences of the United States of America*, 109 (42), 17028-17033. [10.1073/pnas.1212247109](https://doi.org/10.1073/pnas.1212247109)
- Fogelin, R. J. (1994). *Pyrrhonian reflections on knowledge and justification*. Oxford, UK: Oxford University Press.
- (2004). The skeptics are coming! In W. Sinnott-Armstrong (Ed.) *Pyrrhonian skepticism* (pp. 161-173). Oxford, UK: Oxford University Press.
- Friston, K. J. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11 (2), 127-138. [10.1038/nrn2787](https://doi.org/10.1038/nrn2787)
- Gamez, D. (2008). Progress in machine consciousness. *Consciousness and Cognition*, 17 (3), 887-910. [10.1016/j.concog.2007.04.005](https://doi.org/10.1016/j.concog.2007.04.005)

- Geldsetzer, L. (2010). *Nagarjuna: Die Lehre von der Mitte*. Hamburg, GER: Felix Meiner Verlag.
- Gendler, T. S. & Hawthorne, J. (2010). The real guide to fake barns: A catalogue of gifts for your epistemic enemies. In T. S. Gendler (Ed.) *Intuition, imagination, and philosophical methodology* (pp. 98-115). Oxford, UK: Oxford University Press.
- Goldman, A. (2010). Social epistemology. *The Stanford Encyclopedia of Philosophy*.
<http://plato.stanford.edu/archives/sum2010/entries/epistemology-social/>
- Harnad, S. (2007). Ethics of open access to biomedical research: Just a special case of ethics of open access to research. *Philosophy, Ethics, and Humanities in Medicine*, 2 (1), 31-31. [10.1186/1747-5341-2-31](https://doi.org/10.1186/1747-5341-2-31)
- Hart, W., Tullett, A., Shreves, W. & Fetterman, Z. (2015). Fueling doubt and openness: Experiencing the unconscious, constructed nature of perception induces uncertainty and openness to change. *Cognition*, 173, 1-8.
- Hohwy, J. (2013). *The predictive mind*. Oxford, UK: Oxford University Press.
- Holland, O., Knight, R. & Newcombe, R. (2007). A robot-based approach to machine consciousness. In A. Chella & R. Manzotti (Eds.) *Artificial consciousness* (pp. 887-910). Exeter, UK: Imprint Academic.
- Holland, O. & Goodman, R. B. (2003). Robots with internal models: A route to machine consciousness? *Journal of Consciousness Studies*, 10 (4), 77-109.
- Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard University Press.
- (2008). Mental models and deductive reasoning. In J. E. Adler & L. J. Rips (Eds.) *Reasoning: Studies of human inference and its foundations* (pp. 206-222). Cambridge, UK: Cambridge University Press.
- Juengst, E. T. (1998). What does enhancement mean? In E. Parens (Ed.) *Enhancing human traits: Ethical and social implications* (pp. 29-47). Washington, DC: Georgetown University Press.
- Klein, P. (2014). Skepticism. *The Stanford Encyclopedia of Philosophy*.
<http://plato.stanford.edu/archives/sum2014/entries/skepticism/>
- Knauff, M. (2009). Deductive relational reasoning with mental models and visual images. *Spatial Cognition & Computation*, 9 (2), 109-137.
[10.1080/13875860902887605](https://doi.org/10.1080/13875860902887605)
- Knobe, J. (2015). Philosophers are doing something different now: Quantitative data. *Cognition*, 135, 36-38.
- Knobe, J. & Nichols, S. (Eds.) (2008). *Experimental philosophy*. Oxford, UK: Oxford University Press.
- Kuzminski, A. (2008). *How the ancient Greeks reinvented Buddhism*. Lanham, MD: Lexington Books.
- Lambie, J. (2014). *How to be critically open-minded: A psychological and historical analysis*. London, UK: Palgrave Macmillan.
- Landesman, C. (2002). *Skepticism: The central issues*. Oxford, UK: Wiley-Blackwell.
- Landesman, C. & Meeks, R. (2003). *Philosophical skepticism*. Hoboken, NJ: Wiley Blackwell.
- Ludlow, P. (2013). Aaron Swartz was right. *The Chronicle Review*, February 25, 2013.
<http://chronicle.com/article/Aaron-Swartz-Was-Right/137425/>
- Macpherson, F. (2013). The philosophy and psychology of hallucination: An introduction. In F. Macpherson & D. Platchias (Eds.) *Hallucination: Philosophy and psychology* (pp. 1-38). Cambridge, MA: MIT Press.
- Merkel, R., Boer, G., Fegert, J., Galert, T., Hartmann, D., Nuttin, B. & Rosahl, S. (2007). *Intervening in the brain: Changing psyche and society*. Berlin, GER: Springer.
- Metzinger, T. (2000). Introduction: Consciousness research at the end of the twentieth century. In T. Metzinger (Ed.) *Neural correlates of consciousness: Empirical and conceptual questions* (pp. 1-12). Cambridge, MA: MIT Press.
http://mitpress.mit.edu/sites/default/files/titles/content/9780262133708_sch_0001.pdf
- (2003). Phenomenal transparency and cognitive self-reference. *Phenomenology and the Cognitive Sciences*, 2, 353-393.
[10.1023/B:PHEN.0000007366.42918.eb](https://doi.org/10.1023/B:PHEN.0000007366.42918.eb)
- (2004). *Being no one: The self-model theory of subjectivity*. Cambridge, MA: MIT Press.
- (2009). *The ego tunnel. The science of the mind and the myth of the self*. New York, NY: Basic Books.
- (2013a). Why are dreams interesting for philosophers? The example of minimal phenomenal selfhood, plus an agenda for future research. *Frontiers in Psychology*, 4 (476), 1-17. [10.3389/fpsyg.2013.00746](https://doi.org/10.3389/fpsyg.2013.00746)
- (2013b). Two principles of robot ethics. In E. Hilgendorf & J.-P. Günther (Eds.) *Robotik und Gesetzgebung* (pp. 272-286). Baden-Baden, GER: Nomos.
http://www.blogs.uni-mainz.de/fb05philosophieengl/files/2013/07/Metzinger_RG_2013_penultimate.pdf
- (2013c). Spirituality and intellectual honesty. *Mainz: Self-published*. [10.978.300/0415395](https://doi.org/10.978.300/0415395)

- (2013d). The Myth of Cognitive Agency: Subpersonal thinking as a cyclically recurring loss of mental autonomy. *Frontiers in Psychology*, 4 (931), 36-38.
- (2014). First-order embodiment, second-order embodiment, third-order embodiment: From spatiotemporal self-location to minimal phenomenal selfhood. In L. Shapiro (Ed.) *The Routledge Handbook of Embodied Cognition* (pp. 272-286). London, UK: Routledge.
- Metzinger, M. & Hildt, E. (2011). Cognitive enhancement. In J. Illes & B. J. Sahakian (Eds.) *Oxford Handbook of Neuroethics* (pp. 245-264). Oxford, UK: Oxford University Press.
- Metzinger, T. & Windt, J. M. (2014). Die phänomenale Signatur des Wissens: Experimentelle Philosophie des Geistes mit oder ohne Intuitionen? In T. Grundmann, J. Horvath & J. Kipper (Eds.) *Die Experimentelle Philosophie in der Diskussion* (pp. 279-231). Berlin, GER: Suhrkamp.
- Moore, G. E. (1903). The refutation of idealism. *Mind*, 12 (48), 433-453.
- Nielsen, T. A. & Stenstrom, P. (2005). What are the memory sources of dreaming? *Nature*, 437 (7063), 1286-1289. [10.1038/nature04288](https://doi.org/10.1038/nature04288)
- Picard, F. (2013). State of belief, subjective certainty and bliss as a product of cortical dysfunction. *Cortex*, 49 (9), 2494-2500. [10.1016/j.cortex.2013.01.006](https://doi.org/10.1016/j.cortex.2013.01.006)
- Picard, F. & Craig, A. D. (2009). Ecstatic epileptic seizures: a potential window on the neural basis for human self-awareness. *Epilepsy & Behavior*, 16 (3), 539-546.
- Picard, F., Scavarda, D. & Bartolomei, F. (2013). Induction of a sense of bliss by electrical stimulation of the anterior insula. *Cortex*, 49 (10), 2935-2937.
- Picard, F. & Friston, K. (2014). Predictions, perception, and a sense of self. *Neurology*, 83 (12), 1112-1118.
- Planck, M. (1948). *Wissenschaftliche Selbstbiographie. Mit einem Bildnis und der von Max von Laue gehaltenen Traueransprache*. Leipzig, GER: Johann Ambrosius Barth Verlag.
- Popper, K. R. (2013). *The open society and its enemies*. Abington, UK: Routledge.
- Popper, K. R. & Kiesewetter, H. (2003). *Die offene Gesellschaft und ihre Feinde*. Tübingen, GER: Mohr Siebeck.
- Pust, J. (2014). Intuition. *The Stanford Encyclopedia of Philosophy*. <http://plato.stanford.edu/archives/fall2014/entries/intuition/>
- Russell, B. (1912). *The problems of philosophy*. Mineola, NY: Dover Publications.
- (2009). *The basic writings of Bertrand Russell*. Abington, UK: Routledge.
- Ryan, S. (2014). Wisdom. *The Stanford Encyclopedia of Philosophy*. <http://plato.stanford.edu/archives/win2014/entries/wisdom/>
- Sengupta, B., Stemmler, M. B. & Friston, K. J. (2013). Information and efficiency in the nervous system: A synthesis. *PLoS Computational Biology*, 9 (7), e1003157. [10.1371/journal.pcbi.1003157](https://doi.org/10.1371/journal.pcbi.1003157)
- Seth, A. K., Baars, B. J. & Edelman, D. B. (2005). Criteria for consciousness in humans and other mammals. *Consciousness and Cognition*, 14 (1), 119-139. [10.1016/j.concog.2004.08.006](https://doi.org/10.1016/j.concog.2004.08.006)
- Sextus Empiricus, (1987). *Outlines of Pyrrhonism*. Cambridge, MA: Harvard University Press.
- Sinnott-Armstrong, W. (2004). *Pyrrhonian skepticism*. Oxford, UK: Oxford University Press.
- Solomonova, E., Fox, K. C. R. & Nielsen, T. (2014). Methodological considerations for the neurophenomenology of dreaming: A commentary on Windt's "Reporting dream experience". *Frontiers in Human Neuroscience*, 8 (317), 1-3. [10.3389/fnhum.2014.00317](https://doi.org/10.3389/fnhum.2014.00317)
- Stroud, B. (2004). Contemporary Pyrrhonism. In W. Sinnott-Armstrong (Ed.) *Pyrrhonian skepticism* (pp. 174-187). Oxford, UK: Oxford University Press.
- Taylor, R. M. (2014). Open-mindedness: An epistemic virtue motivated by love of truth and understanding. *Philosophy of Education Archive*, 197-205.
- van Noorden, R. (2011). Science publishing: The trouble with retractions. *Nature*, 478 (7367), 26-27. [10.1038/478026a](https://doi.org/10.1038/478026a)
- Windt, J. M. (2013). Reporting dream experience: Why (not) to be skeptical about dream reports. *Frontiers in Human Neuroscience*, 7 (708), 1-15. [10.3389/fnhum.2013.00708](https://doi.org/10.3389/fnhum.2013.00708)
- (2015). *Dreaming*. Cambridge, MA: MIT Press.
- Windt, J. M. & Metzinger, T. (2007). The philosophy of dreaming and self-consciousness: What happens to the experiential subject during the dream state? *The new science of dreaming. Volume 3: Cultural and theoretical perspectives* (pp. 193-248). Westport, CT: Praeger Perspectives.

Beyond Componential Constitution in the Brain

Starburst Amacrine Cells and Enabling Constraints

Michael L. Anderson

Componential mechanism (Craver 2008) is an increasingly influential framework for understanding the norms of good explanation in neuroscience and beyond. Componential mechanism “construes explanation as a matter of decomposing systems into their parts and showing how those parts are organized together in such a way as to exhibit the explanandum phenomenon” (Craver 2008, p. 109). Although this clearly describes some instances of successful explanation, I argue here that as currently formulated the framework is too narrow to capture the full range of good mechanistic explanations in the neurosciences. The centerpiece of this essay is a case study of Starburst Amacrine Cells—a type of motion-sensitive cell in mammalian retina—for which function emerges from structure in a way that appears to violate the conditions specified by componential mechanism as currently conceived. I argue that the case of Starburst Amacrine Cells should move us to replace the notion of mechanistic componential constitution with a more general notion of enabling constraint. Introducing enabling constraints as a conceptual tool will allow us to capture and appropriately characterize a wider class of structure-function relationships in the brain and elsewhere.

Keywords

Componential constitution | Constitution | Constraint | Enabling constraint | Explanation | Functional levels | Levels | Mechanisms | Mechanistic explanation | Neuroscientific explanation | Spatial levels | Starburst amacrine cells | Structure function mapping

1 Introduction

How, in the brain or any other system, does specific function arise from underlying structure? The question is a general one, and also in some sense a vague one, for it asks simultaneously about how structures shape events—generate causes—and also about what kinds of explanations one should aim for in neuroscience. Here I will focus on the second question in the hope of partially illuminating the first. One increasingly influential class of answers to this second question “construes explanation as a matter of decomposing systems into their parts

and showing how those parts are organized together in such a way as to exhibit the explanandum phenomenon” (Craver 2008, p. 109; see also Craver this collection). This is an attractive idea as it is expressed, but what I hope to illustrate here is that the leading formalizations of this general idea (Craver 2008; Craver & Bechtel 2007) place overly restrictive conditions on good mechanistic explanation. In what follows, I lay out the norms of mechanistic explanation, as developed by Craver and Bechtel, and describe some cases that their model nicely cap-

Author

Michael L. Anderson

michael.anderson@fandm.edu

Franklin & Marshall College

Lancaster, PA, U.S.A.

Commentator

Axel Kohler

axelkohler@web.de

Universität Osnabrück

Osnabrück, Germany

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität

Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University

Melbourne, Australia

tures. I then introduce the case of Starburst Amacrine Cells (SACs)—a type of motion-sensitive cell in mammalian retina. In SACs, and in the functionally coupled direction-selective ganglion cells, the function-structure relationship is hard to capture within the Craver/Bechtel mechanistic framework. I argue that we can better capture such cases by replacing the notion of mechanistic componential constitution with the more general notion of enabling constraints.

2 The requirements of mechanistic explanation

Craver (2008) sharply distinguishes between two traditions of understanding scientific explanation: reductive explanation and systems explanation. According to Craver, the first tradition accepts a version of the covering law model of explanation (Hempel 1965) whereby one explains regularities at a given level of organization by showing how these regularities (the laws describing events and their relations) can be derived from theories holding at lower levels. Put differently, one explains a phenomenon of interest by showing how it is to be *expected* based on the laws governing activity at lower levels of organization. This tradition is reductive because when such explanations are successful, one can strictly speaking *do without* the higher-level laws. However convenient they may be for understanding or predicting higher-level phenomena, the higher-level laws do not add, capture, or explain any facts that are not already contained in the lower-level laws. The lower-level laws are scientifically sufficient.

In contrast, in the systems tradition, a phenomenon of interest ψ exhibited by a system S is explained by identifying a set of component parts $\{X\}$ and showing how they are organized such that $S \psi$ s. A systems explanation is similar to reductive explanation in that it too relies on the identification of levels of organization, since it requires identifying the parts of the system S , but, as I note below, it does not aim thereby at the reduction or explanatory absorption of one level by another. Craver & Bechtel write:

In levels of mechanisms, an item X is at a lower level than an item S if and only if X is a component in the mechanism for some activity ψ of S . X is a component in a mechanism if and only if it is one of the entities or activities organized such that $S \psi$'s. For that is what mechanisms are: they are entities and activities organized such that they exhibit a phenomenon. Scientists discover lower levels by decomposing the behavior of a mechanism into the behaviors of its component parts, decomposing the behaviors of the parts into the behaviors of their parts, and so on. (2007, pp. 548–549)¹

As already noted, S is the system that ψ s, or that exhibits phenomenon. It is, for instance, the car (S) that accelerates (ψ), and to explain car acceleration will require identifying the components $\{X\}$ that matter to $S \psi$ -ing. To identify these components and their organization is to explicate the mechanism M that accounts for $S \psi$ -ing. The target of mechanistic explanations of this sort is ψ : “mechanistic explanations are framed by the explanandum phenomenon” (Craver 2008, p. 121) and “[t]he explanandum of a mechanistic explanation is a phenomenon, typically some behavior of a mechanism as a whole” (Craver 2008, p. 139).

In mechanistic explanation, a given X is a component of the mechanism M if and only if it is one of the entities organized such that S exhibits some phenomenon ψ . So the engine, the accelerator, and the gas tank, but not the mudflaps or the windshield wipers are components of M that explain the car accelerating, even though these are *all* parts of the car S . In an

¹ There is a terminological issue that needs to be raised at the outset to avoid confusion. Craver & Bechtel (2007; Craver 2008) usually, but not always, use S to refer to a *mechanism*. In contrast, I will always use S to refer to the *system* or entity exhibiting the explanandum phenomenon ψ , and I introduce the symbol M to refer to the responsible mechanism. I do this because M and S are clearly not identical. Moreover, they *are* (or at least appear to me) to be distinguished in this passage, at least on one reading. I think it is unfortunate that neither Craver nor Bechtel formally and consistently distinguish the system S and the mechanism M in their analysis, for reasons that will become clear at the end of this section. Here I'll attempt to faithfully capture the essence of the Craver–Bechtel mechanistic framework, were it to have included this important distinction.

ideal explanation, the mechanism defined by the parts $\{X\}$ will contain *all* and *only* the components relevant to $S \psi$ -ing (see Craver 2008 for a discussion of constitutive relevance in this context). To identify the parts of M is thus to specify *both* a hierarchical and a functional relationship between M and its parts, and between M and S .

But although mechanistic explanation involves essential reference to hierarchical relationships between levels of organization, it is not thereby a species of *reductive* explanation because in a successful systems explanation nothing is rendered inessential or redundant. The phenomenon ψ is neither *derived* nor *derivable* from laws governing the parts of M ; rather, the parts $\{X\}$ and their relationships simply *are* M , and together explain why $S \psi$ s. The explanatory relationship is not rational derivation, but functional composition: M is physically and functionally *constituted* by its parts, and $S \psi$ s in virtue of that constitution.

Mechanistic explanations are constitutive or componential explanations: they explain the behavior of the mechanism as a whole in terms of the organized activities and interactions of its components. Components are the entities in a mechanism—what are commonly called ‘parts’. (Craver 2008, p. 128)²

Given all this we can add one more criterion for a given X being a part of the mechanism M : each X must be not just a functional but also a *spatial* sub-part of M . As a component of M , X will be at a *lower level* than M , and *smaller than* M : “[b]ecause mechanisms are collections of components and their activities, no component can be larger than the mechanism as a whole, and so levels of mechanisms are ordered by size” (Craver & Bechtel 2007, pp. 549–550). Craver and Bechtel conclude: “[m]ost fundamentally, levels of mechanisms are a species of compositional, or part-whole relations” (Craver & Bechtel 2007, p. 550). In the overall framework developed by Craver

and Bechtel, functional levels and spatial levels generally align.

Thus, although componential mechanistic explanations are not reductive, they generally *are* what one would call “bottom-up”, or perhaps better in this context, “level-restricted”: one explains the phenomenon ψ in S by reference to entities and relations at a lower level of organization, but never the reverse. In componential explanations of this sort, the intrinsic properties of and interactions between the mechanism’s components account for a system’s actions (where “intrinsic” means that such properties—such as the charge of an ion—are either basic to the entity or accounted for by reference to entities and properties at a still lower level of organization). Good mechanistic explanations on this view will not include references to unanalyzed properties of the whole S or M , its “shape” or overall organization, as the relations between the components $\{X\}$ at the lower level will already account for (in fact constitute) these.

This account of mechanistic explanation seems to me a clear and, indeed, compelling model of one kind of explanatory practice in the neurosciences. To satisfy the norms of mechanistic explanation, one must:

1. Identify the phenomenon of interest ψ
2. Identify the system S that ψ s
3. Identify the relevant spatial sub-parts $\{X\}$ of M (and their relevant intrinsic properties)
4. Describe how the parts $\{X\}$ are organized such that $S \psi$ s

At least *prima facie*, a number of instances of successful (albeit incomplete) explanatory models in the neurosciences appear to neatly fit this description. Craver (2008) extensively discusses the mechanistic model of the action potential. Briefly, following the steps above:

1. The phenomenon ψ is the action potential, which consists of the rapid depolarization of neural cells from a resting membrane potential of approximately -70mV toward (and in many cases significantly exceeding) 0mV ; an

² Note that within this framework “componential mechanism”, “constitutive mechanism”, and “compositional mechanism” are synonymous.

equally rapid repolarization; a period of hyperpolarization, where the cell overshoots the normal resting potential; and a gradual return to the resting equilibrium (note that as even this simplified sketch illustrates, ψ will often be in and of itself complex, with many aspects that any adequate model must capture).

2. The system S that ψ s is the neuron.
3. The parts in virtue of which S ψ s include elements of the cell and its surrounding ionic milieu: positively charged K^+ and Na^+ ions; gated, ion-specific membrane channels; and the Na^+/K^+ pump.
4. Finally, the organization that explains ψ includes the following: The resting potential is in fact an equilibrium between two opposing forces: a chemical concentration gradient that pushes Na^+ into the cell and K^+ out of it, and an electrical gradient that pushes K^+ into the cell, each maintained by the selective permeability of the cell to Na^+ and K^+ . Na^+ channels change their conformation in response to current flow (they are voltage-gated) such that they open to allow Na^+ to flow into the cell. As Na^+ flows into the cell this reduces the electrostatic pressure on K^+ , and opens voltage-gated K^+ channels, allowing K^+ to flow out of the cell. The net effect is to push the cell initially toward the electrochemical balance point for Na^+ , which is about +55mV. However, as the membrane potential drops, the Na^+ channels close, thus slowing and eventually stopping the depolarization. The diffusion of K^+ out of the cell combines with the activity of the Na^+/K^+ pump to repolarize the cell, which however overshoots the resting potential due to the fact that the K^+ channels close later than the Na^+ channels, thus allowing K^+ to diffuse out of the cell for an extra millisecond or so during which the cell is hyper-polarized.

Obviously, this remains a sketch (see Craver 2008) or any basic neuroscience textbook for more detail), but it illustrates the main elements of a mechanistic explanation. The intrinsic prop-

erties, actions, and interactions of M 's spatial sub-parts together comprise the mechanism that allows S to ψ and thus explain how S ψ s. One can likewise plausibly sketch the mechanisms that account for spatial long-term memory (e.g., the ability of an animal to return to some location in its environment) in terms of long-term potentiation of synapses in the hippocampus (Craver 2008), although it is worth noting that a more complete account of the functions of hippocampus will have some of the features I describe in 3 and 4 (Buckner 2010; Anderson 2015). Still, the fact that *some* explanations in neuroscience are like this is not under significant dispute.

But this brings us to the question of why I have distinguished M and S in my treatment. Because Craver (2008) does not formally distinguish these, he is never led to ask what the precise relationship is (or could be) between M and S (and between their respective parts). In fact, for Craver the symbol S usually (but not always) refers to what I have been calling M , and he frames his analysis of mechanistic composition entirely in terms of ψ and its mechanism. When he does mention the larger system it is generally to emphasize the fact that not every part of a system S is relevant to the mechanism in virtue of which it ψ s. So what might the committed mechanist say about the relationships between S , M and $\{X\}$? One possibility is: all the parts $\{X\}$ of M will be on a lower level than S . That would be in keeping with the level-restricted character of the framework, and its characteristic alignment between spatial and functional levels. It is certainly a feature of all the examples discussed in its support, including the model of the action potential outlined above. A slightly stronger possibility would be: all the parts $\{X\}$ of M will be spatial sub-parts of S . I don't think anyone would or should endorse this stronger condition, but seeing why will be instructive, and will lead us to the reasons to reject the weaker formulation as well.³

3 On my reading, the framework developed in (Craver 2008) implicitly assumes the weaker condition, although most likely not the stronger one. But for my purposes here it is not crucial to pin this down. If the framework *does* assume the weaker condition, what follows should be read as arguing (contra this model) that there are systems for which functional

The immediate trouble with the stronger formulation is that it collides with a fact noted by Craver (2008), but not otherwise discussed: the mechanism that accounts for S ψ -ing may contain parts that are extrinsic to S (although not to M). For instance, in the mechanism for the action potential, the Na^+ and K^+ ions that are clearly part of M are (at least sometimes) extrinsic to S ; and in embodied accounts of some cognitive processes like mathematics, the mechanism that accounts for a person (P) multiplying (ψ_m -ing) contains parts that are *always* extrinsic to P , such as pencil and paper (Clark 1997; see also [this collection](#)). These entities would arguably *not* be components of the systems that ψ , although they would be components of the mechanisms in virtue of which they ψ . At the very least, this suggests there are some details yet to be worked out about the necessary physical relationships between M and S that implement the hierarchical and functional relationships in virtue of which M can account for S ψ -ing. There will be (presumably rare) cases in which M and S are identical; cases such as the accelerating car where M contains only parts of S ; and cases such as the action potential where M and S cross-cut one another, sharing some but not all of their parts.⁴ There may also turn out to be cases in which they share no parts, perhaps because the parts of M and the parts of S are individuated by different criteria, or because S 's ability to ψ is imposed by or inherited from an entirely extrinsic mechanism (indeed I'll discuss a potential instance of this class of cases later in the paper).

But distinguishing M and S in this way *also* allows one to ask whether all the parts of M need to be at a lower level than S . If not

every X needs to be a spatial sub-part of S , then there is little reason to suppose that each X needs to be on a lower level than S , either. Indeed, I claim that in fact for some systems S the mechanism M will contain items that are neither intrinsic to *nor at a lower level than* S . For instance, I often use other people to help me remember things, in the easiest case by asking them to remind me at some future time. In such a case, this other individual is arguably part of the mechanism responsible for my remembering, but is certainly not for that reason on a lower ontological level than I am, qua remembering system. Moreover, as I will argue when looking at the case discussed below, some relevant parts of M (and certainly M itself) are at a *higher* organizational level than S . Now of course, Craver & Bechtel *define* the concept of lower level in terms of being a part of the mechanism: "an item X is at a lower level than an item S if and only if X is a component in the mechanism for some activity ψ of S " (2007, p. 548). I agree that this holds for the constitutive relationship between *mechanisms* and their parts. But it only holds for all systems S if we assume that all the parts of M are parts of S , and we have seen that this is not always the case. Thus although I think that Craver correctly analyzes the relationship between mechanisms and their parts in terms of constitution, I argue that the more capacious notion of *enabling constraint* better captures the relationship between mechanisms and the systems whose activities they enable.

In any case, with this as background, I now turn to the case of the SAC. In 3, I describe what we know about how the mechanisms in virtue of which the cell operates, and in 4 I discuss the implications of this case for componential mechanistic explanation.

3 Direction selectivity in SAC dendrites: Beyond componential constitution

Starburst Amacrine Cells are axonless neurons found in the retina of mammals and numerous non-mammalian species. Their morphology is planar, with multiple dendrites arrayed, as the

and spatial levels in fact dissociate. If it does not, then what follows should be read simply as offering an account of some of the possible functional relationships between mechanisms and systems, an issue not explored in the original analysis. Either path leads to the same recommended modification of the original model.

⁴ In the case of the action potential, one *might* mount the argument that the system that ψ s is *strictly speaking* $S + \{\text{the nominally non-}S \text{ parts of } M\}$, including the surrounding extracellular fluid. That would make M part of S in this case, but it is not clear to me that this move will be equally attractive in every such case, nor do I think the mechanist is *forced* to adopt this strategy.

name suggests, in a starburst pattern around the cell body (Figure 1).

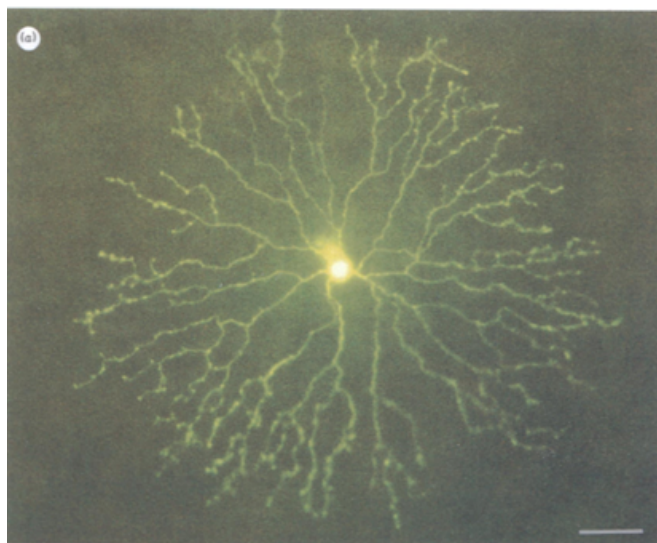


Figure 1: Micrograph of a Starburst Amacrine Cell. Calibration bar 50 μ m. Reprinted from Tauchi & Masland (1984).

SACs form dense, highly overlapping, co-fasciculating layers in the “on” and “off” levels of the inner synaptic layer of the retina, nestled physically and functionally between bipolar cells and direction-selective ganglion cells. Among the most numerous neural cells found in the mammalian retina, they represent a large proportion of the total neural volume in the eye; in the rabbit retina, for example, as much as six meters of SAC dendrites occupy each square millimeter of retinal surface—higher coverage than any other retinal cell by an order of magnitude (Masland 2005; Tauchi & Masland 1984; see Figure 2).

SACs are interesting for multiple reasons. Despite lacking axons, they synthesize and release both excitatory and inhibitory neurotransmitters (ACh (acetylcholine) and GABA (-Aminobutyric acid)) from the distal regions of their dendrites. Both the role and relative proportion of excitatory and inhibitory synaptic connections change over time. Cholinergic synaptic connections between neighboring SACs disappear over development, and GABAergic connections between SACs begin as excitatory but later become inhibitory. However, excitatory cholinergic syn-

apses between SACs and ganglion cells remain (Masland 2005).

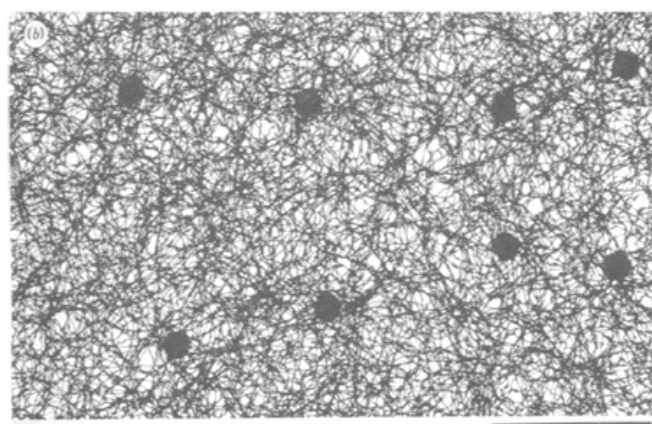


Figure 2: Depiction of the SAC network in peripheral retina. Calibration bar 50 μ m. Reprinted from Tauchi & Masland (1984).

Functionally, SACs play an important role in motion detection, and are part of the overall network for multiple uses including optokinetic eye movement and motion perception (Yoshida et al. 2001). In fact, each dendrite of the SAC acts independently of the others, and signals the presence of stimuli moving centrifugally, that is, from the cell body out in the direction of the signaling dendrite (Euler et al. 2002; see Figure 3). Put differently, each SAC dendrite is a directionally selective spatial sub-part of the overall cell, and this is the functional property that will interest us here. As with so much in the neurosciences, the mechanism that explains this function is complex and not fully understood. It is, however, possible to offer a sketch of it.

As mentioned above, SACs lie between bipolar cells and direction-selective ganglion cells. Bipolar cells thus mediate the initial stimulus such that a moving light causes them to fire in turn as the stimulus moves across the retina. The bipolar cells make excitatory synapses onto the SAC dendrites.⁵ With these basic anatomical facts in view, we can turn to describing

⁵ In fact there are two classes of bipolar cells, “on” and “off”, functionally differentiated by their disposition to respond to stimulus onset vs. stimulus offset—i.e., one responds to light and the other to dark—and anatomically distinguished by whether they synapse onto the “on” or “off” level of the inner synaptic layer (Figure 4). As the mechanisms for direction selectivity in SAC dendrites are the same regardless, I’ll ignore this detail in what follows.

three different aspects of the overall mechanism for direction selectivity: wiring specificity between bipolar cells and the SAC dendrites; lateral inhibition between neighboring SACs; and active elements in the dendrites themselves.

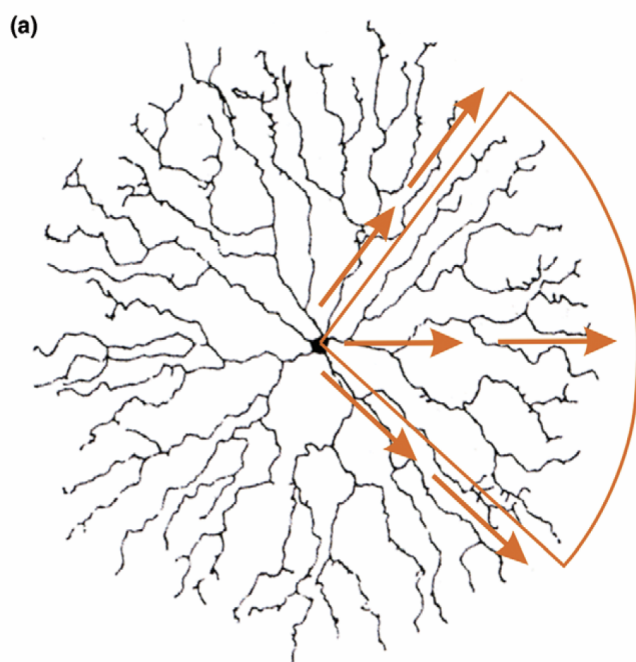


Figure 3: Depiction of direction selectivity in SAC dendrites. Reprinted from Masland (2005).

First, the axonal projections of bipolar cells largely preserve the topography of their inputs, such that neighboring axons come from cells with neighboring inputs, and make neighboring synapses onto post-synaptic cells. What this arrangement means for SACs is that neighboring synapses on the dendrite are likely to come from neighboring bipolar cells, so that when a moving stimulus activates one cell, and then another immediately to its left (say), this will tend to activate a given synapse, and then another immediately to *its* left. Thus, in the case where such a stimulus moves along the direction of a dendritic process, the successive excitatory inputs to that dendrite will tend to reinforce (Demb 2007; Lee & Zhou 2006). This is an important part of the overall mechanism, but is not sufficient by itself to produce the observed directional selectivity, as these inputs would tend to reinforce even during centripetal motion, although this would result in a weaker response at the *distal* process of the dendrite (Hausselt et al. 2007).

Another important part of the mechanism for directional selectivity involves mutual inhibition between neighboring SACs (Figure 5). As a stimulus moves so as to stimulate the centrifugal dendrite of SAC1 (in Figure 5A), reinforcing inputs will cause the release of GABA onto the centripetal dendrite of SAC0, such that even when the light stimuli begins to excite the centripetal dendrite of SAC0, the leading inhibition dominates the signal. Similarly, as the stimulus moves to the *centrifugal* dendrite of SAC0, the successive excitatory inputs from the bipolar cells reinforce, and any inhibitory inputs from the neighboring SAC2 come too late. Moreover, SAC0 will largely inhibit SAC2's response (Figure 5B; Lee & Zhou 2006). An important element of this mechanism involves the relative time-course of ACh and GABA: ACh response from the bipolar cells ramps up and decays fairly quickly, while GABA response is relatively delayed and prolonged (Demb 2007). This temporal asymmetry helps ensure that when inhibition leads it dominates, and vice-versa. The distance between SACs also plays a role. The likelihood of synaptic connections between the distal portion of the dendrites of two SACs—where inhibitory connections are most effective—depends on the distance between the cell bodies. Cells that are very close together or very far apart will thus not mutually inhibit one another (Figure 5C).

Finally, direction selectivity depends upon properties of the dendrite itself. The dendrites are electrically isolated from one another, as a result of both overall cell morphology and the low impedance of the cell body. The uneven distribution of synaptic inputs and outputs also contributes: excitatory inputs from the bipolar cells are distributed along the length of the dendrite, but synaptic outputs are confined to the distal ends (as implied by the two aspects of the overall mechanism described above). A third, active aspect of the local dendritic portion of the mechanism appears to involve voltage-gated calcium channels. These channels lead to amplification of the ACh response beyond what the passive reinforcement caused by successive synaptic transmission from bipolar cells can account for (Hausselt et al. 2007).

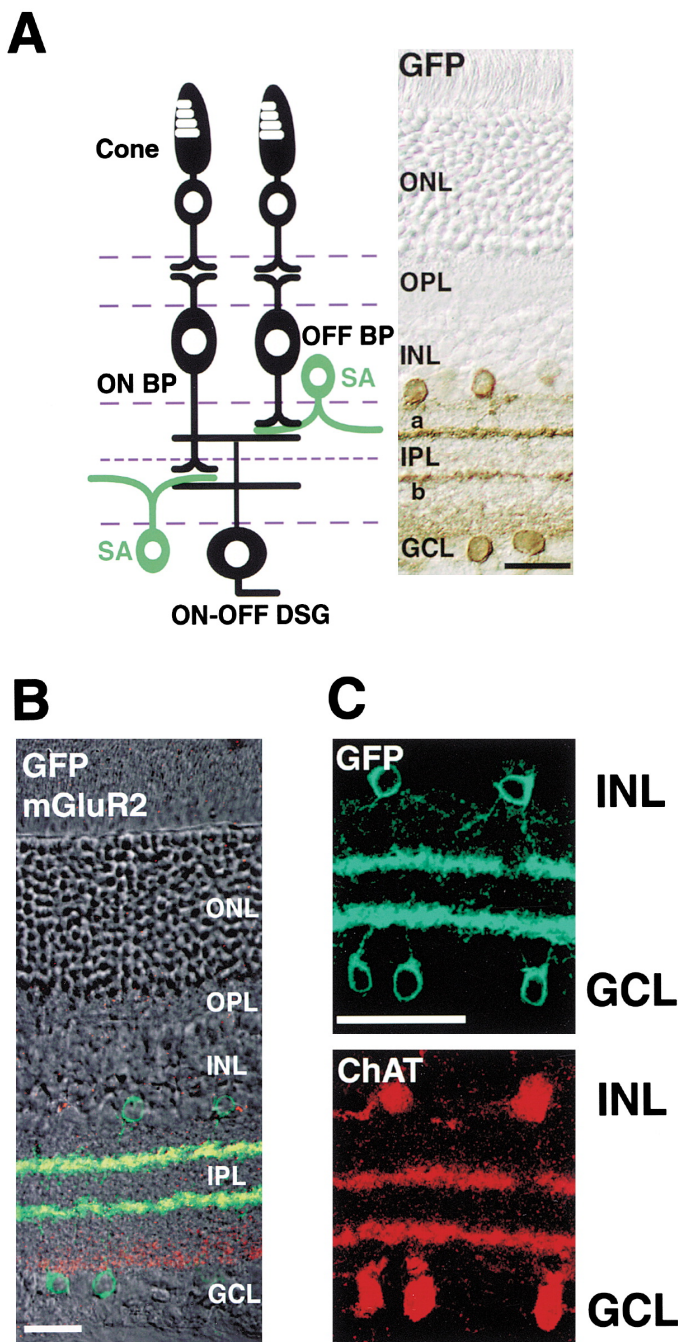


Figure 4: Schematic representation of the layered structure and synaptic relationships between bipolar cells and SACs. Reprinted from Yoshida et al. (2001).

All of these elements combine to produce the direction selectivity of the SAC dendrite. Bipolar cells successively synapse onto the dendritic process, resulting in passive reinforcement of excitatory input that preferentially promotes neurotransmitter release in response to motion in the centrifugal direction. Surrounding SACs selectively inhibit centripetal excitation,

as a result of the different temporal activation profiles of GABA and ACh; the asymmetric distribution of input and output synapses; and the relative spatial placement of the SACs. And voltage-gated calcium channels in the dendrite actively amplify the centrifugal signal. Although this sketch leaves out many of the known details, and there remain many details still to be worked out, I believe it is sufficient to warrant the conclusion that this is (a) an instance of mechanistic explanation that (b) does not have the level-restricted character of the (canonical) mechanistic explanations laid out above. I spell out the reasons for this conclusion in the next section.

4 Constitution and constraint

We can most readily see why this case represents an interesting challenge for componential mechanism by fitting it to the four steps outlined in section 2, above.

1. Identify the phenomenon of interest ψ
2. Identify the system S that ψ s
3. Identify the relevant spatial sub-parts $\{X\}$ of M (and their relevant intrinsic properties)
4. Describe how the parts $\{X\}$ are organized such that $S \psi$ s

The specific phenomenon of interest ψ_{ds} is direction selectivity or, more precisely, the release of neurotransmitter in and only in response to motion in a specific centrifugal direction. The system S_{ds} that exhibits ψ_{ds} is the dendrite of the SAC. It is also easy to say what the parts $\{X_{ds}\}$ of the mechanism M_{ds} are in virtue of which the dendrite ψ_{ds} , and how they are organized. I have provided that sketch above. Finally, it seems right to say, following Craver (2008), that the relationship between M_{ds} and its parts $\{X_{ds}\}$ is one of componential constitution, such that all the parts $\{X_{ds}\}$ are at a lower level than M_{ds} , and together constitute M_{ds} . But now it gets interesting for componential mechanistic explanation as currently developed. For only some of the parts of M_{ds} —including the voltage gated calcium channels, and the input and output synapses—are at a lower (spatial) level than the

dendrite S_{ds} . The inhibitory dendrites of the neighboring SACs are at the same level as S_{ds} , the bipolar cells and their spatial relations are arguably at a higher level than S_{ds} (although one might wish to screen these off as mere *inputs* to the mechanism), and the mechanism M as a whole in virtue of which S_{ds} ψ_{ds} s is *certainly* at a higher level than, and is in no way a physical or functional component of S_{ds} .

I think this example demonstrates that not every mechanistic explanation will have the “bottom-up” or “level-restricted” character that the mechanism for the action potential has, where function is built entirely from the capacities of lower-level components and their interactions. In the SAC dendrite, we appear to have a case *not* of a system that ψ s in virtue of the capacities and relations of its components (and that could in turn be thought of as a component supporting the activities of a larger functional system), but rather very nearly the reverse: a system that ψ s in virtue of the properties of and interactions in the higher-level system of which it is a part. That is, the SAC dendrite is not functionally related to its surrounds as a component to a higher-level system; nor is the higher-level system related to the SAC dendrite as one of *its* components. Instead, I want to say that the higher-level mechanism M acts as an *enabling constraint* on S .

Before providing a bit more in the way of substantial analysis of the concept of an enabling constraint, let us pause to consider one way in which a supporter of componential mechanistic explanation might resist this conclusion by redefining the system S_{ds} to include the mechanism M_{ds} . I think this is not a viable option for a number of reasons. First, it would appear to violate standard usage: neuroscientists speak of direction-selective dendrites, and *not* of a directionally selective network spanning several retinal layers. The debate in the neuroscientific literature concerns *not* the definition of the direction-selective system, but the relative role of intrinsic and extrinsic mechanisms for dendritic direction selectivity in SACs (Hausselt et al. 2007; Lee & Zhou 2006).

Second, it appears that the mechanism as a whole is *not* direction selective. Any given

SAC, for instance, and certainly the network as a whole, signals motion in *all* directions. Even if we restrict the definition of M_{ds} to the entities in virtue of which *one* particular SAC dendrite is directionally selective, the symmetry of the mechanism—the fact that SACs *mutually constrain* one another and the same bipolar cells synapse onto more than one SAC dendrite—strongly suggests that *very same mechanism* generates right direction selectivity in the rightward-reaching dendrite in SAC0, and left direction selectivity in the leftward-reaching dendrite in SAC2 (e.g., in Figure 4). The mechanism, that is, does not have the same direction selectivity as either of the dendrites. Rather, it’s as if when you turn the crank one way (i.e., the stimulus moves one way) the mechanism produces one output; and when you turn it the other way, it produces the other output.

This suggests a different way to illustrate the limitations of componential mechanism as formulated. Craver writes that the explanandum phenomenon ψ is “typically some behavior of the mechanism as a whole” (Craver 2008, p. 139), and he thus might insist, contra my way of formulating his framework in 2, that it is the mechanism M and not the system S that exhibits ψ . In this case, because I have agreed that the parts $\{X\}$ in fact constitute M , any conflict between functional and spatial levels disappears. But in the case before us it seems that the mechanism *responsible* for, say, rightward direction selectivity does not in fact *exhibit* rightward direction selectivity. So the functional puzzle reasserts itself in a different guise.⁶

One might nevertheless insist on distinguishing these mechanisms in subtle ways—perhaps M_{ds0} includes these synapses from bipolar cells, but not those synapses, while M_{ds2} includes those synapses but not these. I doubt whether this can work, because explaining direction selectivity in *either* direction will require reference to the excitatory inputs from bipolar cells to the centrifugal dendrite, and the inhibitory inputs from the overlapping centripetal dendrite, which are in turn a result of the excitatory inputs from the *very same* bipolar cells synapsing

⁶ Thanks to an anonymous reviewer for pointing out this way of expressing the matter.

onto the centrifugal dendrite. But let us take the possibility as granted. Then one seems forced to say something along the following lines: the mechanism as a whole ψ s, but *signals* ψ -ing with the dendrite.

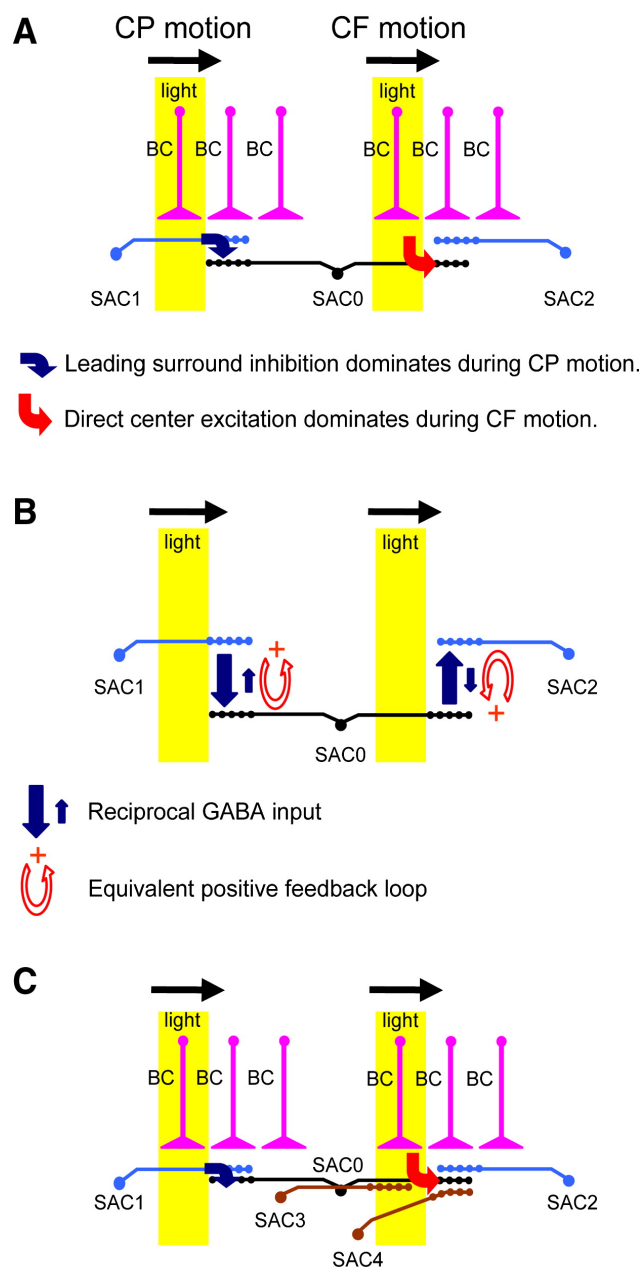


Figure 5: Lateral inhibition between neighbouring SACs contributes to direction selectivity in the dendrites. Reprinted from [Lee & Zhou \(2006\)](#).

Let us consider this possibility carefully. As I intimated above, scientists debate the relative importance of intrinsic and extrinsic mechanisms for dendritic selectivity in SACs. [Hausselt et al. \(2007\)](#) note that direction se-

lectivity in SAC dendrites persists in the presence of GABA and glycine receptor antagonists, which would deactivate the portions of the normal mechanism that involve mutual inhibition between neighboring SACs. In these circumstances, one might argue that *only* the portions of the original mechanism *intrinsic* to the dendrite matter in the explanation of direction selectivity, and in such a case it is clearly the dendrite that ψ s. What shall we say, then, when we remove the antagonists from the system and reapply the same directional stimulus, resulting in neurotransmitter release from this dendrite? One option is: whereas before the dendrite ψ 'd, now it merely signals the ψ -ing of the larger mechanism. But it seems clear to me that, if the dendrite can ψ , then adding network interactions that *aid and enhance* (that is, do not in any sense prevent) ψ -ing can hardly cause it to *not* ψ , but only signal ψ . This points to a fourth and final reason to reject the general move to extend the neural system S to include the mechanism M whenever it is (or contains entities that are) on a higher level than S : one would apparently need the ability to rigorously distinguish between ψ -ing and *signaling* ψ in an overall system where to ψ is generally also to signal it—that is, where signaling and doing are deeply intertwined. Thus, I believe we must insist: the dendrite ψ s.

For all these reasons, I do not think it is wise to hold onto level-restricted explanations and componential composition by fiat. Instead, it is time to expand the scope of mechanistic explanation by considering the various ways in which systems S relate to the mechanisms M that enable their activities. I think the case of SACs is especially important because it illustrates one way in which local selectivity in parts of a network can be the result of the interplay of excitation and mutual inhibition between non-selective parts of that network, which is clearly something that we need to understand better if we are to accurately characterize the functional mechanisms at work in both small and large-scale brain networks ([Anderson et al. 2013](#)). But other structure-function relationships appear to call equally for a broader account of mechanistic explanation. For instance,

the direction-selective ganglion cell DSGC (Direction-Selective Ganglion Cell), mentioned briefly above, responds to stimuli moving only in its preferred direction (which of course varies cell-to-cell). In this case, there do not appear to be *any* intrinsic mechanisms for the direction selectivity of the DSGC. Rather, SAC dendrites selectively synapse onto DSGCs with preferred stimuli antiparallel to the SAC dendrite preference (Briggman 2011) thus suppressing responses to motion in the non-preferred direction. DSGCs seem to simply *inherit* their selectivity via their synaptic contact with SACs—and, in fact, elimination of SACs from the retina abolishes direction selectivity in DSGCs (Yoshida et al. 2001). Here I just don't see any case for a compositional relationship between the mechanism (or its parts) and the selective system. Instead, the relevant mechanism synapses onto the relevant system, and by suppressing a sub-set of its response tendencies, induces selectivity.

This brings us finally back to the notion of “constraint”, which I think may help us understand the full range of mechanism/system relationships in the brain. The term constraint has been used in myriad ways in the literature on scientific explanation. In evolutionary biology, scientists refer for instance to stability constraints (Schlosser 2007) and both universal and local developmental constraints on evolvability (Maynard Smith et al. 1985). There are also law-like constraints on the possible states of physical systems generally (Lange 2011). None of these capture the sense of “constraint” that will be most helpful to us here.

One notion that gets us close is the idea of a “capacity constraint”, that is, a limitation on the capacity of a process that might take the form of changing the relative probabilities of the range of possible process outcomes (Sansom 2009). This certainly has the right flavor, for in the mechanism under discussion above it appears that the excitatory and inhibitory interactions between bipolar cells and neighboring SACs bias the outcome of the dendritic processing of the moving stimulus. But insofar as a capacity constraint is generally conceptualized in terms of the reduction

of some pre-existing whole ability—in Sansom's (2009) example, being handcuffed limits one's ability to move one's hands—this does not offer quite the right organizing frame for explanation in neuroscience.

The reason is that in the neurosciences we want to understand not just the capacities of entities, but how the structured interactions between entities give rise to *functions*, which are, crucially, *differential* and *differentiating* processes (that is, they differ from one another, and they differentiate between stimuli). Capacities in the sense of general powers (the capacity to generate an action potential, say) are necessary conditions for functions, but they are not yet functions; the DSGC is strictly speaking *non-functional* in the absence of SACs, even though it will continue to exercise its capacity to fire action potentials in response to inputs from bipolar cells. Constraints of the sort under investigation here serve to limit capacities, but in so doing they enable functions; they result in an *enhancement* (not a reduction) of the abilities of the system (and the organism).

For this reason I propose to analyze the general functional (and, crucially, *non-hierarchical*) relationship between mechanisms and systems in the following way: an *enabling constraint* is a relationship between entities and/or mechanisms at a particular level of description and a functional system at the same or a different level, such that the entities/mechanisms bias (i.e., change the relative probabilities of) the outcomes of processing by the system. Such enabling constraints offer necessary but not sufficient conditions for the instantiation of differential function in neural systems. Because enabling constraints are synchronic rather than diachronic, the idea shares the same explanatory advantage that the relation of constitution has over the relation of “causation” (when understood, e.g., as an event involving the transmission of some property, power, or conserved quantity from one entity to another). As Craver & Bechtel (2007) point out, such a conception of causation does not accommodate interlevel functional relationships well, because these are often synchronic and symmetric, whereas causa-

tion of this sort is temporal and asymmetric.⁷ In addition, enabling constraints can be *mutual*, which gives the idea an advantage over both causation and constitution as an analysis of functional relationships in the brain.

Enabling constraint =_{Df} A physical relationship between a functional system S and entities $\{X\}$ (and/or mechanism M), at the same or different level of description, such that $\{X\}$ (and/or M) changes the relative probabilities of various possible functional outcomes of activity in S .

To understand function not just in systems like SAC dendrites and DSGCs, but also in the large scale networks that are partially constituted by the Transiently Assembled Local Neural Subsystems TALoNS (Transiently Assembled Local Neural Subsystems) crucial to the functioning of a dynamic brain (Anderson 2015), we need to accept that there is a broader range of relationships that mechanisms can have to functional systems, beyond componential constitution. Function in TALoNS results not from structured interactions between stable, autonomous low-level components, but rather from the interplay between the capacities of lower-level entities and higher-level network dynamics. That interplay, I argue, is best analyzed in terms of the mutual constraint that exists between bottom-up and top-down, feed-forward and feed-back mechanisms in the brain.

5 Conclusion

Although mechanistic explanation as developed by Craver & Bechtel (2007; Craver 2008) does seem to accurately characterize one kind of explanation in neuroscience, and one kind of func-

tional arrangement in neural systems, I've argued here that the formulation is not wide enough to capture the variety of mechanisms in the brain. When we formally distinguish the system S from the mechanism M in virtue of which S exhibits the explanandum phenomenon Ψ , we see that although it seems correct to describe the relationship between M and its parts $\{X\}$ in terms of constitution, it will only sometimes be the case that S is (partially) constituted by $\{X\}$.

As an alternative to the relationship of componential constitution, I have offered the notion of an *enabling constraint* that can exist between a system and the mechanism(s) in virtue of which it has its various functions. SAC dendrites appear to have their function in virtue of the enabling constraints imposed by entities at the same and higher levels of organization; and DSGC function is enabled by the constraints imposed by the SAC dendrites. In neither case is it appropriate to describe the relationship between the mechanism M and the relevant system S in terms of constitution, nor are all (or, in the case of DSGCs arguably any) of the parts $\{X\}$ of M components of S .

Overall, I hope to have made the case that moving beyond level-restricted mechanistic explanation will allow us to better capture the variety of neural systems that emerge from the constant, constraining, biasing interplay between feed-forward, feedback, bottom-up, and top-down processes in the dynamic brain.

⁷ For instance, what explains why a neuron has a particular functional property cannot be an event involving the transmission of some property, power or conserved quantity from the parts of the neuron to the whole, because if causes must precede their effects, this would appear require that there be a time prior to which the neuron did not have the functional property conferred by its parts. Interlevel functional relationships do not generally appear to be temporal in this way. Rather, for Craver and Bechtel, what explains the functional property of the neuron is the way it is *constituted* by its parts. Enabling constraints are also synchronic in the relevant way, and so the view I am advocating here is also able to accommodate such cases of interlevel functional relationships.

References

- Anderson, M. L. (2015). *After phrenology: Neural reuse and the interactive brain*. Cambridge, MA: MIT Press.
- Anderson, M. L., Kinnison, J. & Pessoa, L. (2013). Describing functional diversity of brain regions and brain networks. *NeuroImage*, 73, 50-58. [10.1016/j.neuroimage.2013.01.071](https://doi.org/10.1016/j.neuroimage.2013.01.071)
- Briggman, K. L., Helmstaedter, M. & Denk, W. (2011). Wiring specificity in the direction-selectivity circuit of the retina. *Nature*, 471 (7337), 183-188. [10.1038/nature09818](https://doi.org/10.1038/nature09818)
- Buckner, R. L. (2010). The role of the hippocampus in prediction and imagination. *Annual Review of Psychology*, 61, 27-48. [10.1146/annurev.psych.60.110707.163508](https://doi.org/10.1146/annurev.psych.60.110707.163508)
- Clark, A. (1997). *Being there: Putting brain, body and world together again*. Cambridge, MA: MIT Press.
- (Ed.) (2015). Embodied prediction. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Craver, C. F. (2008). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford, UK: Oxford University Press.
- (Ed.) (2015). Levels. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Craver, C. F. & Bechtel, W. (2007). Top-down causation without top-down causes. *Biology & Philosophy*, 22 (4), 547-563. [10.1007/s10539-006-9028-8](https://doi.org/10.1007/s10539-006-9028-8)
- Demb, J. B. (2007). Cellular mechanisms for direction selectivity in the retina. *Neuron*, 55 (2), 179-186. [10.1016/j.neuron.2007.07.001](https://doi.org/10.1016/j.neuron.2007.07.001)
- Euler, T., Detwiler, P. B. & Denk, W. (2002). Directionally selective calcium signals in dendrites of starburst amacrine cells. *Nature*, 418 (6900), 845-852. [10.1038/nature00931](https://doi.org/10.1038/nature00931)
- Hauselt, S. E., Euler, T., Detwiler, P. B. & Denk, W. (2007). A dendrite-autonomous mechanism for direction selectivity in retinal starburst amacrine cells. *PLoS Biology*, 5 (7), e185. [10.1371/journal.pbio.0050185](https://doi.org/10.1371/journal.pbio.0050185)
- Hempel, C. G. (1965). *Aspects of scientific explanation*. New York, NY: Free Press.
- Lange, M. (2011). Conservation laws in scientific explanations: Constraints or coincidences? *Philosophy of Science*, 78 (3), 333-352. [10.1086/660299](https://doi.org/10.1086/660299)
- Lee, S. & Zhou, Z. J. (2006). The synaptic mechanism of direction selectivity in distal processes of starburst amacrine cells. *Neuron*, 51 (6), 787-799. [10.1016/j.neuron.2006.08.007](https://doi.org/10.1016/j.neuron.2006.08.007)
- Masland, R. H. (2005). The many roles of starburst amacrine cells. *Trends in Neurosciences*, 28 (8), 395-396. [10.1016/j.tins.2005.06.002](https://doi.org/10.1016/j.tins.2005.06.002)
- Maynard Smith, J., Burian, R., Kauffman, S., Alberch, P., Campbell, J., Goodwin, B., Lande, R., Raup, D. & Wolpert, L. (1985). Developmental constraints and evolution. *Quarterly Review of Biology*, 60 (3), 265-287.
- Sansom, R. (2009). The nature of developmental constraints and the difference maker argument for externalism. *Biology & Philosophy*, 24 (4), 441-59. [10.1007/s10539-008-9121-2](https://doi.org/10.1007/s10539-008-9121-2)
- Schlosser, G. (2007). Functional and developmental constraints on life cycle evolution: An attempt on the architecture of constraints. In R. Sansom & R. Brandon (Eds.) *Integrating evolution and development: from theory to practice* (pp. 113-173). Cambridge, MA: MIT Press.
- Tauchi, M. & Masland, R. H. (1984). The shape and arrangement of the cholinergic neurons in the rabbit retina. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 223 (1230), 101-119. [10.1098/rspb.1984.0085](https://doi.org/10.1098/rspb.1984.0085)
- Yoshida, K., Watanabe, D., Ishikane, H., Tachibana, M., Pastan, I. & Nakanishi, S. (2001). A key role of starburst amacrine cells in originating retinal directional selectivity and optokinetic eye movement. *Neuron*, 30 (3), 771-780. [10.1016/S0896-6273\(01\)00316-6](https://doi.org/10.1016/S0896-6273(01)00316-6)

Carving the Brain at its Joints

A Commentary on Michael L. Anderson

Axel Kohler

When neuroscientists explain the biological basis of a phenomenon of interest, they usually try to identify the parts of a system that seem to do the relevant job, and propose a model of how those parts interact to produce the phenomenon. This mechanistic framework of explanation is widely used and has been investigated from a philosophical point of view by different authors. In his target article, Michael Anderson poses a challenge to the currently dominant version of mechanistic explanation as advocated, e.g., by Carl Craver. Taking empirical results and explanatory models from studies on retinal starburst amacrine cells as a starting point, Anderson suggests that the current framework for mechanistic explanation should be extended to include a differentiation between systems and mechanisms, which would allow more leeway in understanding processing in the nervous system. Mechanisms can then be seen to provide enabling constraints on the functioning of systems, where the mechanisms do not need to be subsumed under the system and do not even have to be on the same organizational level. Although Anderson's proposal is interesting and worth exploring, I am unconvinced that this extension conforms to real-world explanatory practice and/or is necessary for accommodating the understanding of direction-selectivity in the retina. I examine another sample of research on starburst amacrine cells, where the integration of empirical data and computational models shows that, on close inspection, it is distributed networks to which certain characteristics are ascribed—a situation that can be handled with the available tools of mechanistic explanation.

Keywords

Constitution | Direction selectivity | Enabling constraint | Enabling constraints | Mechanism | Mechanistic explanation | Motion processing | MT | Neuroscience | Neuroscientific explanation | Starburst amacrine cells | Top-down causation | V1

1 Introduction

One of the dominant frameworks of explanatory practice in the neurosciences and the biological sciences in general is the model of mechanistic explanation proposed in its modern form by Bechtel & Richardson (1993) and recently extended by Carl Craver (2007). Mechanistic explanations describe entities and activities that together bring about a phenomenon of interest (Machamer et al. 2000). When we are interested in how vision works, for example, we try to localize the relevant parts of the brain, and identify components and their types of interactions in order to understand how we can see things (Bechtel 2008). This model of mechan-

istic explanation is thought to capture the dominant explanatory practice in the biological sciences (Bechtel & Richardson 1993), but normative claims are also made with respect to the adequacy of explanatory accounts. Craver (2007) proposes a number of constraints on constitutive mechanistic explanation in order to decide whether a mechanistic model is viable or not.

In his target article, Michael Anderson (this collection) takes current models of mechanistic explanation as a starting point for proposing an important extension of the existing accounts. In previous models, the system that

Commentator

Axel Kohler

axelkohler@web.de
Universität Osnabrück
Osnabrück, Germany

Target Author

Michael L. Anderson

michael.anderson@fandm.edu
Franklin & Marshall College
Lancaster, PA, U.S.A.

Editors

Thomas Metzinger

metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu
Monash University
Melbourne, Australia

exhibits a phenomenon and the mechanism that explains the phenomenon were not separated. Sometimes parts of the system can be screened off with respect to the phenomenon at hand. The windshields of a car and its radio components are not really important in order to understand how it drives, for example. It's fine to say that the whole car drives, but that only the relevant components (engine, axles, tires) are doing the mechanistic work. Focusing on the essential components of a mechanism within a larger system is unproblematic. But Anderson worries about more complex cases in the neurosciences where the system displaying a phenomenon does not encompass the relevant mechanism producing the phenomenon and might not even be on the same level of description as the mechanistic components.

Anderson wants to demonstrate that componential constitution is not sufficient as a model of mechanistic explanation for the processing of directional selectivity in the retina. Mechanisms computing direction of motion are already available at the earliest stages of the visual hierarchy. The vital components of direction selectivity in the retina could be identified. In particular, in recent discussion starburst amacrine cells (SAC) have been viewed as a mechanistic substrate of motion processing. The SACs receive input from bipolar cells, which are not themselves directionally selective, and provide output to direction-selective ganglion cells (Zhou & Lee 2008). The SACs themselves seem to be the core component for retinal motion selectivity (Park et al. 2014; Yoshida et al. 2001).

Examining the current models of how direction selectivity is created in SACs, Anderson takes note of a discrepancy between how direction selectivity is mechanistically achieved and to which parts it is ascribed. He argues for a distinction between the system S that Ψ s (that is, exhibits direction selectivity) and the mechanism M that accounts for S 's Ψ -ing. For the case at hand, the SACs themselves or even just single dendritic compartments of SACs Ψ , but a much broader network of neighboring SACs and bipolar cells needs to be considered in order to provide a mechanistic account of SAC direction

selectivity. Anderson proposes this distinction as an important extension of Craver and Bechtel's model of mechanistic explanation. This has two major advantages, according to Anderson: (1) there can be entities and actions that play a role for M , but are not necessarily parts of S . This allows a certain flexibility in defining the system that displays Ψ , while at the same time including all relevant components in the mechanistic account of S 's Ψ -ing. (2) But if there are parts of M that don't need to be spatially subsumed under S , neither do they need to be at a lower level than S . So even the requirement of componential constitution might be relaxed to allow for higher-level mechanistic components that play an important role in S 's Ψ -ing.

As an alternative account of the relationship between mechanisms M and the respective systems S , Anderson proposes that M acts as an enabling constraint on S :

[A]n enabling constraint is a relationship between entities and/or mechanisms at a particular level of description and a functional system at the same or a different level, such that the entities/mechanisms bias (i.e., change the relative probabilities of) the outcomes of processing by the system. (this collection, p. 12)

In the case of retinal direction selectivity, the mechanistic interaction between neighboring SACs and BCs acts as an enabling constraint for the direction selectivity of a specific SAC dendritic compartment (i.e., the system).

The most straightforward move by proponents of existing models of mechanistic explanation, as Anderson (this collection) also notes, would be to claim that the differentiation of system and mechanism is vacuous. Only the mechanism as a whole can do the work. Even in complex cases, one just has to pick out the right subparts of the network (specific synapses, specific compartments of neurons) that together produce the phenomenon of interest. Anderson provides a number of arguments against this way of extending the concept of mechanism/system, which I would like to briefly summarize:

1. Neuroscientists just don't talk about complex directionally selective networks, but about the direction selectivity of certain dendritic branches.
2. The mechanism as a whole does not display a specific direction selectivity (it is not rightward-selective etc.), it only contributes to the specific selectivity in the respective SAC dendrites. The mechanism contributes to different kinds of selectivities in different dendrites.
3. Making fine-grained distinctions between subparts (synapses, axon branches, dendrites etc.) of the very same neurons that contribute to different directional selectivities is implausible.
4. When the whole network is said to be direction-selective (i.e., it Ψ s), what about the dendrite itself? Is it supposed to only signal direction selectivity (signal Ψ -ing)? It is unlikely that a clear distinction between Ψ -ing and signaling Ψ -ing can be made.

The aim of this commentary is twofold. First, I would like to argue that the described cases can be handled by current models of mechanistic explanation when one considers the options of reconstituting the phenomena and top-down causation. Second, using another example of research on SACs, I would like to show that the straightforward ascription of direction selectivity to the SAC dendrites is at least debatable. When looking at how empirical results are often integrated with computational models of direction selectivity, it becomes clear that those phenomena can only be understood by considering the distributed nature of the involved networks.

2 Reconstituting the phenomena and top-down causation

Anderson proposes a separation between systems and mechanisms. No matter whether the system is constrained to be a dendritic compartment or whether it is extended to encompass all mechanistically relevant parts, there are tools available to describe the respective situation. The mechanistic model does not necessarily consider systems in isolation from the environ-

ment or surrounding processes. Even if the system is defined as the dendrite only, factors influencing dendritic processing as well as the embedding of the system in the overall economy, its organization, have to be considered in order to arrive at an understanding of the system's functioning (Bechtel 2008, pp. 148–150). On the other hand, I would like to argue that we have good reason to extend the boundaries of the system to encompass all the contributing parts. This is a situation in which the original ascription of a function to a system part has to be revised to accommodate new findings. This process is termed *reconstituting the phenomena* by Bechtel & Richardson (1993). Although direction selectivity was thought to be bound to or even intrinsically generated in SAC dendrites, it turns out that the system can only be understood in combination with other neural elements that vitally contribute to the mechanism in question.

One advantage that Anderson suggests comes with the differentiation of system and mechanism is that mechanistic components can then be set at a different level of organization than the relevant system. The SAC dendrite is at a lower level compared to the input from bipolar cells and the network structure (bipolar cells and neighboring SACs) that enables SAC direction-selectivity. But once the question of how exactly we should carve up the brain into systems and mechanisms has been answered, I don't think that complex inter-level relationships are much of an issue for mechanistic accounts. They can be easily accommodated within the framework of top-down causation proposed by Craver & Bechtel (2007). They suggest that any reference to inter-level interactions can be analyzed in terms of within-level causal relationships between parts of entities, where parts and entities are related in a constitutive fashion and entities can be located on different levels. Emphasizing the fact that complex inter-level interactions often need to be considered in order to offer adequate explanatory accounts in neuroscience is important, but it is not outside the scope of current models of mechanistic explanation.

3 Systems and mechanisms for direction selectivity

Since the processing of direction selectivity in the retina is currently a very active research field, there is substantial controversy concerning the relevant entities and activities that contribute to the mechanism, as Anderson points out in his target article. Some accounts focus on local processes within the SAC dendrites themselves (Hausselt et al. 2007), while others draw a broader picture of a multi-component process, where the exact arrangement of cell types and their compartments is vital for direction selectivity (Lee & Zhou 2006). For our purposes here, I would like to use a most recent update on SAC function offered by the group working with Sebastian Seung. The group uses high-resolution electron-microscopy images of brain tissue to reconstruct complete brain networks on a cellular level. Apart from trained reconstruction experts, the project also makes use of so-called “citizen neuroscientists”—volunteers who contribute to the reconstruction process through an online platform that employs gaming features to guide and motivate the community effort (<http://www.eyewire.org>).

In their study, Seung and colleagues used images from the mouse retina to analyze SAC circuitry. They took a closer look at the exact wiring between bipolar cells (BCs) and SACs (Kim et al. 2014). BCs provide input to SACs, but do not show any directional selectivity by themselves. The main point of the article is to show that different BC subtypes display different patterns of connectivity with SACs. By analyzing branch depth and contact area, they could show that one subtype (BC2) has mainly connections close to the soma, while another subtype (BC3a) has more connections far from the soma in the outer parts of the dendrites. Importantly, the BC subtypes, in turn, have different intrinsic visual response latencies. BC2 seems to lag BC3a by 50ms and more. It can be shown that the differential connectivity patterns and the divergent latencies add up to produce selectivity for a preferred direction of movement going out from the soma on the respective dendrite in accordance with empirical results.

What is important about the paper is not just the main result itself. Any empirical observation may be overruled in the (near) future. So it is not particularly relevant whether these exact cell types and this exact type of wiring is vital for the phenomenon at hand. What I found intriguing in this study, however, was how the relevant mechanism was described and how the data were integrated with a computational model of direction selectivity, reflecting a recent trend in the neurosciences to combine biological and computational perspectives in explanatory accounts. It shows how neuroscientists pick out the relevant parts of a system that contribute to a specific phenomenon in question. The proposed computational model (Fig. 1a; Kim et al. 2014) maps the biological entities onto specific parts of the computational circuit. The output element at the lower part of the figure is the SAC. The input stems from BC2 (left) and BC3a (right); their respective response properties are captured as delay values and sustained vs. transient response types. The circuit combines elements of classical models of direction selectivity, the Reichardt (Fig. 1b) and the Barlow-Levick detectors (Fig. 1c). Clearly, the direction selectivity cannot be attributed to any one of the system components in isolation. Mechanistic accounts and the corresponding computational models both point to the whole complex of entities as the relevant system that achieves directional selectivity.

In its computational abstraction, the model can be thought of as a canonical system of directional selectivity. Similar models have also been applied to different hierarchical levels of neural processing and different species. For example, mechanisms of directional selectivity have been studied for a long time in the fly visual system. With very different neural elements and wiring, a system of interconnected neurons achieves directional selectivity with response properties closely resembling the Reichardt-type of motion detector (Borst & Euler 2011). Again, only the combination of elements from different processing stages succeeds in delivering direction selectivity as a system. On a cortical level, direction selectivity has been first described for complex cells of the

primary visual cortex (V1) in the seminal work of David Hubel & Torsten Wiesel (1962). Without offering a quantitative computational model, they nevertheless suggest a hypothetical connectivity pattern between different cell types that might underlie the observed responses to moving patterns in complex cells (Hubel & Wiesel 1962, Fig. 20). The model shares features with other motion detectors; a mapping between components is possible.

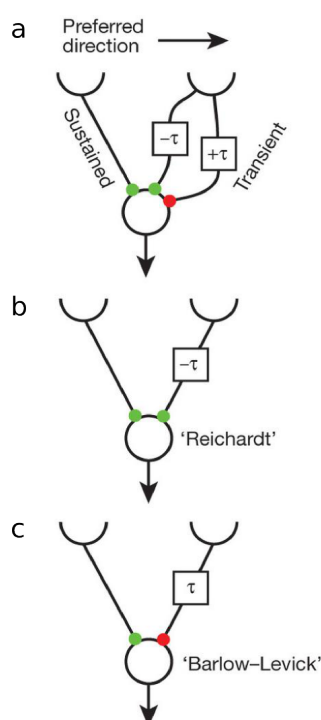


Figure 1: Computational models of direction selectivity (a) The selectivity of SACs described in Kim et al. (2014) can be modeled with a computational framework using a combination of sustained and transient response properties as well as excitatory and inhibitory connections. The displayed wiring would lead to direction selectivity for rightwards motion. The proposed model can be considered to combine previous classical models of direction selectivity, the Reichardt detector (b) and the Barlow-Levick model (c). Green dots indicate excitatory and red dots inhibitory synapses. $-\tau$ indicates a temporal lead and $+\tau$ a temporal lag. Reprinted by permission from Macmillan Publishers Ltd: Nature (Kim et al. 2014), copyright (2014).

When it comes to motion selectivity in the brain, one of the most intensively studied cortical areas is the middle temporal (MT) region.

The region was first described in the macaque (Dubner & Zeki 1971; Zeki 1974) and owl monkeys (Allman & Kaas 1971). The human homolog, the human MT complex (hMT+; Tootell et al. 1995; Zeki et al. 1991), turned out to be a collection of areas with related response properties (Amano et al. 2009; Kolster et al. 2010). Again, to understand the direction selectivity of MT, it is necessary to consider the cooperation of cells in MT and the input processing stages, mainly from V1. This cooperation and the need for an integrated perspective is emphasized in empirical studies (Saproo & Serences 2014) as well as computational models of MT functioning (Rust et al. 2006). Only the V1-MT system as a whole is understood to deliver motion selectivity as output of the MT stage.

But in terms of the role of MT in motion processing, a case could be made in support of Anderson's suggested distinction between a system that exhibits a certain selectivity and the mechanism that produces this selectivity. The apparent locality and modularity of motion processing in MT is based on very selective deficits in patients with lesions in and around MT (Zeki 1991; Zihl et al. 1983). And stimulation of MT with transcranial magnetic stimulation (TMS) in healthy participants leads to selective deficits in motion perception (Beckers & Hömberg 1992; Beckers & Zeki 1995; Hotson et al. 1994; Sack et al. 2006). In a recent study, patients undergoing brain surgery near MT could be investigated with electrical stimulation (Becker et al. 2013). Only stimulation of MT and a related area nearby, MST, led to an inability to perform a simple motion-detection task, a rather specific result concerning the relevance. Results of that kind drive the intuition that the system that is responsible for motion perception, independent of any cortical areas that might mechanistically contribute to the processing chain leading up to MT (like V1), are localized in MT.

Lesion and other interference studies (e.g., with TMS) are suggestive, but there are also well-known difficulties with interpreting the results. Lesions mostly affect larger parts of the brain and are rarely limited to a single cortical site. As such it is often hard to identify the actual parts of the complex brain networks that

are affected. The advantage of stimulation techniques is that the interference is temporary and can be precisely targeted on a specific location. But, given the rich connectivity structure of neural networks, stimulation effects can be seen even in remote target sites (Bestmann et al. 2004; Sack et al. 2007). In addition, TMS studies have shown that activity of MT might not even be sufficient for conscious motion perception without the involvement of V1 (Pascual-Leone & Walsh 2001; Silvanto et al. 2005). There are also further empirical as well as philosophical reasons for rejecting the claim that motion perception can be attributed to MT in a stringent fashion (Madary 2013), which I won't discuss here.¹

So while at first glance MT is a very strong candidate for straightforward and very local attribution of function, it seems again that the relevant system is more appropriately described on a network level. The tendency to see system parts as vital for a function may also stem from the limitations of our employed methods. Lesion cases and interference techniques are commonly interpreted as being informative about the relevant gray-matter structures that are affected by the lesion or stimulation. But there is evidence that interference with white-matter connections between network parts can be even more incapacitating than gray-matter damage. It has long been known that frontoparietal areas are implicated in a deficit of visuospatial attention called *neglect*. But very recently Thiebaut de Schotten et al. (2005, 2011) revealed that the properties of fiber connections between frontal and parietal sites are most predictive of visuospatial processing capacities, and that their electrical stimulation leads to severe deficits. Transferring this insight to the case of MT, we simply have most direct access to the cortical gray-matter centers involved in motion processing, and since they are vital components of the system, this also leads to

corresponding deficits when they are affected or stimulated. But this might conceal the fact that motion selectivity is a product of a wider network that crucially depends on integrated processing for proper functioning.

In sum, I think that close inspection of how direction selectivity is investigated and treated in neuroscientific research is in disagreement with Anderson's arguments (1) and (3). Although it is true that investigators sometimes refer loosely to local elements as displaying a certain characteristic, the corresponding detailed and extended accounts of direction selectivity give credit to the distributed nature of the relevant systems that figure in explanations. Even considering the case of conscious motion perception, it is unclear whether the presumed locality of motion representation stands up to stringent tests. Rather, it seems to be a case of localized interference with a distributed system where damage to vital hubs leads to fundamental deficits.

4 Conclusion

In this commentary, I have defended the claim that the current tools of mechanistic explanation are sufficient for accommodating the explanatory goals in current neuroscience, particularly in the special case of direction selectivity in the retina and other neural systems. A closer look at explanatory practice shows that, in representative cases of empirical research, models of direction selectivity have to take a number of components in a distributed network into account in order to provide a full-fledged description of the relevant processes. On the philosophical side, the conceptual tools of "reconstituting the phenomena" (Bechtel & Richardson 1993) and "top-down causation" (Craver & Bechtel 2007), offered by existing models of mechanistic explanation, might be sufficient for capturing the problematic cases to which Anderson ([this collection](#)) points.

On the other hand, Anderson's proposal ([this collection](#)) to extend existing models of mechanistic explanation with the notion of enabling constraints is very interesting and might offer an avenue to more nuanced mechanistic

¹ Madary (2013) uses two sets of empirical results to show that representation of motion cannot be ascribed to MT simpliciter. One is the recent emphasis on spontaneous activity making significant contributions to the state of sensory systems—they add content referring to the attentional or sensorimotor state of the organism to input-derived sensory representations. The other demonstrates that in MT specifically, the response properties of cells can be quite variable and are not consistently related to perceptual content only.

descriptions of systems in their contextual embedding. In almost all relevant cases in neuroscience research, there are various external factors influencing the workings of a system, and it is often difficult to draw clear boundaries between vital and non-vital, but nevertheless highly influential system components. Anderson's framework would offer a viable solution for handling those modulatory constraints. Resolving this debate will also depend on a clear conception of how the entities that display a certain phenomenon are best identified and described.

References

- Allman, J. M. & Kaas, J. H. (1971). A representation of the visual field in the caudal third of the middle temporal gyrus of the owl monkey (*Aotus trivirgatus*). *Brain Research*, 31 (1), 85-105. [10.1016/0006-8993\(71\)90635-4](https://doi.org/10.1016/0006-8993(71)90635-4)
- Amano, K., Wandell, B. A. & Dumoulin, S. O. (2009). Visual field maps, population receptive field sizes, and visual field coverage in the human MT+ complex. *Journal of Neurophysiology*, 102 (5), 2704-2718. [10.1152/jn.00102.2009](https://doi.org/10.1152/jn.00102.2009)
- Anderson, M. L. (2015). Beyond componential constitution in the brain: Starburst Amacrine Cells and enabling constraints. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Bechtel, W. (2008). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. New York, NY: Taylor & Francis.
- Bechtel, W. & Richardson, R. C. (1993). *Discovering complexity: Decomposition and localization as scientific research strategies*. Princeton, NJ: Princeton University Press.
- Becker, H. G. T., Haarmeier, T., Tatagiba, M. & Gharabaghi, A. (2013). Electrical stimulation of the human homolog of the medial superior temporal area induces visual motion blindness. *Journal of Neuroscience*, 33 (46), 18288-18297. [10.1523/JNEUROSCI.0556-13.2013](https://doi.org/10.1523/JNEUROSCI.0556-13.2013)
- Beckers, G. & Hömberg, V. (1992). Cerebral visual motion blindness: Transitory akinetopsia induced by transcranial magnetic stimulation of human area V5. *Proceedings of the Royal Society of London, Series B, Biological Sciences*, 249 (1325), 173-178. [10.1098/rspb.1992.0100](https://doi.org/10.1098/rspb.1992.0100)
- Beckers, G. & Zeki, S. (1995). The consequences of inactivating areas V1 and V5 on visual motion perception. *Brain*, 118 (1), 49-60. [10.1093/brain/118.1.49](https://doi.org/10.1093/brain/118.1.49)
- Bestmann, S., Baudewig, J., Siebner, H. R., Rothwell, J. C. & Frahm, J. (2004). Functional MRI of the immediate impact of transcranial magnetic stimulation on cortical and subcortical motor circuits. *European Journal of Neuroscience*, 19 (7), 1950-1962. [10.1111/j.1460-9568.2004.03277.x](https://doi.org/10.1111/j.1460-9568.2004.03277.x)
- Borst, A. & Euler, T. (2011). Seeing things in motion: Models, circuits, and mechanisms. *Neuron*, 71 (6), 974-994. [10.1016/j.neuron.2011.08.031](https://doi.org/10.1016/j.neuron.2011.08.031)
- Craver, C. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford, UK: Oxford University Press.
- Craver, C. & Bechtel, W. (2007). Top-down causation without top-down causes. *Biology and Philosophy*, 22 (4), 547-563. [10.1007/s10539-006-9028-8](https://doi.org/10.1007/s10539-006-9028-8)
- Dubner, R. & Zeki, S. M. (1971). Response properties and receptive fields of cells in an anatomically defined region of the superior temporal sulcus in the monkey. *Brain Research*, 35 (2), 528-532. [10.1016/0006-8993\(71\)90494-X](https://doi.org/10.1016/0006-8993(71)90494-X)
- Hausselt, S. E., Euler, T., Detwiler, P. B. & Denk, W. (2007). A dendrite-autonomous mechanism for direction selectivity in retinal starburst amacrine cells. *PLoS Biology*, 5 (7), e185. [10.1371/journal.pbio.0050185](https://doi.org/10.1371/journal.pbio.0050185)
- Hotson, J. R., Braun, D., Herzberg, W. & Boman, D. (1994). Transcranial magnetic stimulation of extrastriate cortex degrades human motion direction discrimination. *Vision Research*, 34 (16), 2115-2123. [10.1016/0042-6989\(94\)90321-2](https://doi.org/10.1016/0042-6989(94)90321-2)
- Hubel, D. H. & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160, 106-154.
- Kim, J. S., Greene, M. J., Zlateski, A., Lee, K., Richardson, M., Turaga, S. C., Purcaro, M., Balkam, M., Robinson, A., Behabadi, B. F., Campos, M., Denk, W., Seung, H. S. & EyeWisers, (2014). Space-time wiring specificity supports direction selectivity in the retina. *Nature*, 509 (7500), 331-336. [10.1038/nature13240](https://doi.org/10.1038/nature13240)
- Kolster, H., Peeters, R. & Orban, G. A. (2010). The retinotopic organization of the human middle temporal area MT/V5 and its cortical neighbors. *Journal of Neuroscience*, 30 (29), 9801-9820. [10.1523/JNEUROSCI.2069-10.2010](https://doi.org/10.1523/JNEUROSCI.2069-10.2010)
- Lee, S. & Zhou, Z. J. (2006). The synaptic mechanism of direction selectivity in distal processes of Starburst Amacrine Cells. *Neuron*, 51 (6), 787-799. [10.1016/j.neuron.2006.08.007](https://doi.org/10.1016/j.neuron.2006.08.007)
- Machamer, P., Darden, L. & Craver, C. (2000). Thinking about mechanisms. *Journal of Philosophy*, 67, 1-25.

- Madary, M. (2013). Placing area MT in context. *Journal of Consciousness Studies*, 20 (5-6), 93-104.
- Park, S. J. H., Kim, I.-J., Looger, L. L., Demb, J. B. & Borghuis, B. G. (2014). Excitatory synaptic inputs to mouse on-off direction-selective retinal ganglion cells lack direction tuning. *Journal of Neuroscience*, 34 (11), 3976-3981. [10.1523/JNEUROSCI.5017-13.2014](https://doi.org/10.1523/JNEUROSCI.5017-13.2014)
- Pascual-Leone, A. & Walsh, V. (2001). Fast backprojections from the motion to the primary visual area necessary for visual awareness. *Science*, 292 (5516), 510-512. [10.1126/science.1057099](https://doi.org/10.1126/science.1057099)
- Rust, N. C., Mante, V., Simoncelli, E. P. & Movshon, J. A. (2006). How MT cells analyze the motion of visual patterns. *Nature Neuroscience*, 9 (11), 1421-1431. [10.1038/nn1786](https://doi.org/10.1038/nn1786)
- Sack, A. T., Kohler, A., Linden, D. E., Goebel, R. & Muckli, L. (2006). The temporal characteristics of motion processing in hMT/V5+: combining fMRI and neuronavigated TMS. *NeuroImage*, 29 (4), 1326-1335. [10.1016/j.neuroimage.2005.08.027](https://doi.org/10.1016/j.neuroimage.2005.08.027)
- Sack, A. T., Kohler, A., Bestmann, S., Linden, D. E., Dechent, P., Goebel, R. & Baudewig, J. (2007). Imaging the brain activity changes underlying impaired visuospatial judgments: simultaneous fMRI, TMS, and behavioral studies. *Cerebral Cortex*, 17 (12), 2841-2852. [10.1093/cercor/bhm013](https://doi.org/10.1093/cercor/bhm013)
- Saproo, S. & Serences, J. T. (2014). Attention improves transfer of motion information between V1 and MT. *Journal of Neuroscience*, 34 (10), 3586-3596. [10.1523/JNEUROSCI.3484-13.2014](https://doi.org/10.1523/JNEUROSCI.3484-13.2014)
- Silvanto, J., Cowey, A., Lavie, N. & Walsh, V. (2005). Striate cortex (V1) activity gates awareness of motion. *Nature Neuroscience*, 8 (2), 143-144. [10.1038/nn1379](https://doi.org/10.1038/nn1379)
- Thiebaut de Schotten, M., Urbanski, M., Duffau, H., Volle, E., Levy, R., Dubois, B. & Bartolomeo, P. (2005). Direct evidence for a parietal-frontal pathway subserving spatial awareness in humans. *Science*, 309 (5744), 2226-2228. [10.1126/science.1116251](https://doi.org/10.1126/science.1116251)
- Thiebaut de Schotten, M., Dell'Acqua, F., Forkel, S. J., Simmons, A., Vergani, F., Murphy, D. G. & Catani, M. (2011). A lateralized brain network for visuospatial attention. *Nature Neuroscience*, 14 (10), 1245-1246. [10.1038/nn.2905](https://doi.org/10.1038/nn.2905)
- Tootell, R. B., Reppas, J. B., Kwong, K. K., Malach, R., Born, R. T., Brady, T. J. & Belliveau, J. W. (1995). Functional analysis of human MT and related visual cortical areas using magnetic resonance imaging. *Journal of Neuroscience*, 15 (4), 3215-3230.
- Yoshida, K., Watanabe, D., Ishikane, H., Tachibana, M., Pastan, I. & Nakanishi, S. (2001). A key role of starburst amacrine cells in originating retinal directional selectivity and optokinetic eye movement. *Neuron*, 30 (3), 771-780. [10.1016/S0896-6273\(01\)00316-6](https://doi.org/10.1016/S0896-6273(01)00316-6)
- Zeki, S. M. (1974). Functional organization of a visual area in the posterior bank of the superior temporal sulcus of the rhesus monkey. *Journal of Physiology*, 236 (3), 549-573.
- (1991). Cerebral akinetopsia (visual motion blindness): A review. *Brain*, 114, 811-824. [10.1093/brain/114.2.811](https://doi.org/10.1093/brain/114.2.811)
- Zeki, S., Watson, J. D. G., Lueck, C. J., Friston, K. J., Kennard, C. & Frackowiak, R. S. J. (1991). A direct demonstration of functional specialization in human visual cortex. *Journal of Neuroscience*, 11 (3), 641-649.
- Zhou, Z. J. & Lee, S. (2008). Synaptic physiology of direction selectivity in the retina. *Journal of Physiology*, 586 (Pt 18), 4371-4376. [10.1113/jphysiol.2008.159020](https://doi.org/10.1113/jphysiol.2008.159020)
- Zihl, J., von Cramon, D. & Mai, N. (1983). Selective disturbance of movement vision after bilateral brain damage. *Brain*, 106 (2), 313-340. [10.1093/brain/106.2.313](https://doi.org/10.1093/brain/106.2.313)

Functional Attributions and Functional Architecture

A Reply to Axel Kohler

Michael L. Anderson

In his commentary ([Kohler this collection](#)) on my target article ([Anderson this collection](#)), Axel Kohler suggests that componential mechanism ([Craver 2008](#)) in fact suffices as a framework for understanding function-structure relationships, even in complex cases such as direction selectivity in Starburst Amacrine Cells. Here I'll argue that while Kohler is correct that the framework *can* accommodate such cases, this approach misses an opportunity to draw important distinctions between what appear to be different sorts of relationships between functioning systems and the mechanisms in virtue of which they function. I tentatively suggest further that the avenue that one prefers may turn on whether one expects the functional architecture of the brain to be primarily componential and hierarchical ([Craver 2008](#); [this collection](#)) or typically more complex than that ([Pessoa 2014](#)).

Keywords

Constitution | Direction-selective ganglion cells | Enabling constraint | Explanation | Hierarchy | Levels | Mechanisms | Mechanistic explanation | Neuroscientific explanation | Starburst amacrine cells | Structure function mapping

Author

[Michael L. Anderson](#)
michael.anderson@fandm.edu
Franklin & Marshall College
Lancaster, PA, U.S.A.

Commentator

[Axel Kohler](#)
axelkohler@web.de
Universität Osnabrück
Osnabrück, Germany

Editors

[Thomas Metzinger](#)
metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Introduction

In my target article ([Anderson this collection](#)), I argued that the complexity of the function-structure relationships that give rise to direction selectivity in Direction-Selective Ganglion Cells (DSGCs) and in the dendrites of Starburst Amacrine Cells (SACs) represent a challenge to componential mechanism as currently formulated ([Craver 2008](#)). First, I argued that distinguishing between the system *S* that exhibits the target phenomenon ψ , and the

mechanism *M* in virtue of which it ψ s allows one to paint a more nuanced picture of the various ways entities can be organized so as to give rise to observed function. Second, I suggested that the function-structure relationships in these particular cases appeared to violate the bottom-up hierarchical assumptions at the center of the componential mechanistic framework, which requires that the components of *M* in virtue of which a system exhibits ψ are at a lower level of organization than *S*. In the cases under

discussion, I argued that some parts of the mechanism in virtue of which SAC dendrites function are at a *higher* level of organization than the dendrite, and that parts of the mechanism in virtue of which DSGCs function are at the *same* level. Moreover, I noted that in neither of these cases were all the entities that constituted M constitutive parts (components) of S.

To accommodate such cases, I recommended extending the notion of mechanistic *constitution* with the notion of an *enabling constraint*: mechanisms, we should say, enable function in systems by changing the relative probabilities of functional outcomes of activity in S. I suggested that this change would allow us to more accurately characterize the variety of structure–function relationships in the brain (and in other complex systems). However, in his commentary on my article ([Kohler this collection](#)), Axel Kohler argues that such an extension is unnecessary, for in fact the componential mechanistic framework of Craver and Bechtel ([Craver 2008](#); [Craver & Bechtel 2007](#)) can accommodate these cases.

Kohler is correct. The extension is strictly speaking unnecessary, and componential mechanistic explanation can offer one plausible characterization of function–structure relationships in these cases. In fact, it is probably the case that *no* example or set of examples *ever* forces one to give up on an explanatory framework (certainly not one as well-motivated and useful as componential mechanism). What examples such as these *can* do, however, is illuminate the potential *advantages* of a new approach, and I would like to use the opportunity offered by this reply to reiterate what I take some of those advantages to be.

2 Three possible system-mechanism relationships

In my target article ([Anderson this collection](#)) I suggested that once one distinguishes between the system S that ψ s and the mechanism M in virtue of which it does so, it is easy to see that there are three possible relationships between M and S. First, the components of M can all also

be components of S, such that M is a relevant sub-component of S. Let's call this relationship R1. A relationship of type R1 obtains between the drive-train of an automobile and the automobile as a whole. Second, (R2), M and S can be identical. I can't think of an uncontroversial example of this relationship, and imagine that such a case is relatively rare. Third and finally, (R3), M and S can cross-cut in various ways, sharing some but not all of their parts. In my view, for instance, it is the neuron the fires an action potential, but not all of the entities that comprise the mechanism for generating action potentials are also part of the neuron. For example, the ions in the extracellular fluid that are crucial for establishing the membrane potential are not part of the neuron, although they are clearly part of the mechanism. Similarly, I argued in my target article that in the case of direction-selectivity in SAC dendrites, although it is the dendrite itself that is directionally selective, many of the parts of the relevant mechanism are not in fact parts of the dendrite. Moreover, in the case of DSGCs, the cell and the mechanism in virtue of which it is direction-selective share at most *one* part: the synapse between the SAC dendrite and the DSGC.

One advantage of making these distinctions, I believe, is that it allows one to see quite clearly when top-down constraints are responsible for function, as I argued is the case for direction selectivity in SAC dendrites. But Kohler suggests that appearances may be misleading here. In fact, he argues, we should “reconstitute the phenomenon” by recognizing that the relevant direction-selective system is *not* the SAC dendrite, but is rather the dendrite + the non-dendritic elements of the mechanism, including other SACs. This larger system can be then be treated within the standard framework of componential mechanism. We can call this approach to addressing these sorts of cases “the Kohler strategy”.

As I noted in my target article, the Kohler strategy is certainly open to the mechanist. It does, however, have the following effects. First, it tends to make the systems of the brain to which functions are attributed relatively *larger* and more diffuse, which arguably reduces preci-

sion. Second, it would in effect turn all apparent instances of R3 into instances of R2.¹ I noted above that I thought the class of R2 would be small. If I am right about the prevalence of R3 functional relationships in the brain, then this strategy would make R2 very large. But it would do so essentially by legislation, as a way of preserving the universal applicability of the componential mechanist framework. How forced this appears will depend on how closely one believes the guiding assumptions of that framework match the architectural facts of the brain. We will return to this last point after reviewing some of the considerations that appear to favor the Kohler strategy.

3 Motivations for the Kohler strategy

Kohler maintains that actual scientific practice in fact supports the Kohler strategy. Exhibit A in his argument is a recent article (Kim et al. 2014) detailing part of the mechanism for visual motion detection. Kohler reproduces a figure depicting their model, and argues that the inclusion of the distributed network in the model suggests that the authors are strictly speaking attributing function to the whole system as depicted:

Although it is true that investigators sometimes refer loosely to local elements as displaying a certain characteristic, the corresponding detailed and extended accounts of direction selectivity give credit to the distributed nature of the relevant systems that figure in explanations. (Kohler this collection, p. 6)

I agree that this is one possible interpretation of the practice. But here is another: these scientists are distinguishing between the system that exhibits the phenomenon and the mechanism that produces it, and are open to different sorts

of relationships between them. Consider the following from the paper Kohler discusses:

Research on [the visual detection of motion] has converged upon the SAC. An SAC dendrite is more strongly activated by motion outward from the cell body to the tip of the dendrite, than by motion in the opposite direction. *Therefore an SAC dendrite exhibits DS*, and outward motion is said to be its ‘preferred direction’. Note that it is incorrect to assign a single such direction to a SAC, because each of the cell’s dendrites has its own preferred direction. DS persists after blocking inhibitory synaptic transmission, when the only remaining inputs to SACs are BCs, which are excitatory. As the SAC exhibits DS but its BC inputs exhibit little or none, *DS appears to emerge from the BC–SAC circuit*. (Kim et al. 2014, p. 331; emphases added)

Far from seeming loose, the attribution of direction-selectivity to the dendrite appears to me clear and precise. Moreover, note that in the final sentence quoted above, the attribution of direction-selectivity to the cell is reinforced, even in the context of a reference to the mechanism as the “BC-SAC circuit”. Indeed, I would argue it is natural and permissible to gloss the last clause in the following way: “DS *in the dendrite* appears to emerge from the BC-SAC circuit.” On this reading, of course, the authors of this article would be proposing an R3 functional relationship such that parts of the mechanism are on a higher level of organization than the system exhibiting the phenomenon.

That these authors are open to R3 relationships of various sorts appears to be reinforced by a line later in the paper:

Previous research suggests that On–Off direction-selective ganglion cells *inherit their DS from SAC inputs* owing to a strong violation of Peters’ rule. (Kim et al. 2014, p. 335; emphasis added)

Here again we see the same pattern: a clear attribution of direction-selectivity *to the DSGC* in

¹ Actually, there are some questions here, for there seem to be *obvious* instances of R3 with which the mechanist is and should be entirely comfortable, e.g., the neuron and the mechanism of the action potential. So presumably this strategy would be employed *only* when the relationship appeared to violate the “lower-level entity” constraint. I’ve not the space to pursue this further here, so will note only that *selectively* pursuing the Kohler strategy would need separate justification.

the same sentence as a reference to the distal mechanism (the SACs), in the context of what is obviously an R3 relationship between system and mechanism. Thus, while I agree that the Kohler strategy is viable, I don't see that consideration of scientific practice forces us to adopt it, or even necessarily favors it.

So what might be other reasons for adopting the Kohler strategy over extending mechanism to include enabling constraints? As I mentioned at the end of the previous section, the matter might come down to how closely one thinks the architectural facts about the brain match the guiding assumptions of the componential mechanistic framework. If one expects that the brain is at root a decomposable or nearly-decomposable system of well-defined interacting components, then componential mechanism does indeed seem like a very appropriate framework for capturing at least the majority of its functional relationships (with the few exceptions to be dealt with perhaps as secondary elaborations or special cases). If, however, one takes seriously the notion that the brain is a massive network marked by multiple, nested, cross-cutting, dynamic hierarchies interacting in bottom-up, top-down, feed-forward and feedback fashions (Pessoa 2014), then one might wish for some of the explanatory flexibility that the notion of enabling constraints appears to offer. I, of course, am in this latter camp (Anderson 2015).

4 Conclusion

As Kohler correctly points out, it is possible to accommodate these complex cases of function-structure relationships within the componential mechanistic framework, by reconstituting the phenomenon and ascribing function to the whole mechanism that produces it. I have tried to indicate what I think some of the costs are to the Kohler strategy, including an apparent conflation of R2 and R3 functional relationships and a potential loss of grain in our ascriptions of function to structure. For some, paying these costs will be preferable to the proposed alternative, which might appear to require the admission of spooky top-down causes into our ontology.

For those who instead want to maintain the greater attributional specificity that appears to conform to scientific discourse, and in the current case to explain direction selectivity *in the SAC dendrite*, then I would argue that the most promising strategy is to recognize the ways in which functional parts (including networks) can impose constraints on other functional parts, at whatever relative level of organization. Adopting this strategy will of course focus attention on the nature of these constraints, whether bottom-up, top-down, or synpedionic. I would hope that the careful study of such R3 relationships as those showcased here would result in a better understanding of the varieties of causal interactions in complex systems.

References

- Anderson, M. L. (2015). *After phrenology: Neural reuse and the interactive brain*. Cambridge, MA: MIT Press.
- (2015). Beyond componential constitution in the brain: The case of starburst amacrine cells. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Craver, C. F. (2008). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford, UK: Oxford University Press.
- (2015). Levels. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Craver, C. F. & Bechtel, W. (2007). Top-down causation without top-down causes. *Biology and Philosophy*, 22 (4), 547-563. [10.1007/s10539-006-9028-8](https://doi.org/10.1007/s10539-006-9028-8)
- Kim, J. S., Greene, M. J., Zlateski, A., Kisuk, L., Richardson, M., Turaga, S. C., Purcaro, M., Balkam, M., Robinson, A., Behabadi, B. F., Campos, M., Denk, W., Seung, H. S. & the EyeWriters, (2014). Space-time wiring specificity supports direction selectivity in the retina. *Nature*, 509 (7500), 331-336. [10.1038/nature13240](https://doi.org/10.1038/nature13240)
- Kohler, A. (2015). Carving the brain at its joints: A commentary on Michael L. Anderson. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Pessoa, L. (2014). Understanding brain networks and brain organization. *Physics of Life Reviews*, 11 (3), 400-435. [10.1016/j.plrev.2014.03.005](https://doi.org/10.1016/j.plrev.2014.03.005)

What a Theory of Knowledge–How Should Explain

A Framework for Practical Knowledge beyond Intellectualism and Anti-Intellectualism

Andreas Bartels & Mark May

We argue against both intellectualist and anti-intellectualist approaches to knowledge-how. Whereas intellectualist approaches are right in denying that knowledge-how can be convincingly demarcated from knowledge-that by its supposed non-propositional nature (as is assumed by the anti-intellectualists), they fail to provide positive accounts of the obvious phenomenological and empirical peculiarities that make knowledge-how distinct from knowledge-that. In contrast to the intellectualist position, we provide a minimal notion of conceptuality as an alternative demarcation criterion. We suggest that conceptuality gives a sound basis for a theory of knowledge-how which is empirically fruitful and suitable for further empirical research. We give support to this suggestion by showing that, by means of an adequate notion of conceptuality, five central peculiarities of knowledge-how as compared to knowledge-that can be accounted for. These peculiarities are its context-bound, impenetrable and implicit nature, as well as the automatic and continuous forms of processing that are connected to it.

Keywords

(anti-)intellectualism | (non-)propositionality | Conceptuality | Disposition(ality) | Intuitive knowledge | Knowledge representation | Knowledge-how | Knowledge-that | Practical mode of thinking | Sensorimotor knowledge

Authors

[Andreas Bartels](#)

andreas.bartels@uni-bonn.de
Rheinische Friedrich-Wilhelms-Universität
Bonn, Germany

[Mark May](#)

mm@hsu-hh.de
Helmut-Schmidt-Universität
Hamburg, Germany

Commentator

[Ramiro Glauer](#)

ramiro.glauer@ovgu.de
Otto-von-Guericke-Universität
Magdeburg, Germany

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Introduction

In this paper, we shall argue against both intellectualist and anti-intellectualist approaches to knowledge-how,¹ for their failing to provide a suitable framework for empirical research on the subject of practical knowledge. Anti-intellectualists propose, following Ryle (1949), that intelligent action embodies “practical knowledge”, which is distinguished from “theoretical knowledge” by its manifesting abilities or dispositions. Intellectualists, in contrast, claim that there is only one sort of knowledge that is characterized by having propositional content (e.g., Stanley 2011b). Practical knowledge, according to intellectualists, is rather distinguished by how propositional contents are *applied* in action. Whereas intellectualist approaches (e.g., Stanley 2011b), we shall argue, are right in denying that practical knowledge can be convincingly demarcated from theoretical knowledge by its supposed non-propositional nature, nevertheless they fail to provide a conceptual framework in which the peculiarities by which practical knowledge stands out could be made visible.

On the other hand, anti-intellectualists (e.g., Newen & Jung 2011) often present phenomenologically-motivated identifications of forms of practical knowledge with certain representational formats. Classificatory schemas without theoretical foundation—that is, without a general conceptual framework within which these classifications naturally emerge, and without any clear-cut specification of the explanatory tasks that have to be fulfilled by that classification—have only limited value as a manual for empirical research. Such schemas cannot even be judged according to explanatory productivity or completeness.

The first part of the paper (sections 2, 3, and 4) will be concerned with the shortcomings of both intellectualist and anti-intellectualist approaches, partly programmed by Ryle’s famous, but also somewhat misleading, exposition of the subject. The perception of these deficiencies of both intellectualist and anti-intellectual-

ist approaches leads us to the conclusion that a philosophical framework for practical knowledge, in order to provide a basis for further empirical research, has in the first instance to lay some firm meta-theoretical ground.

The second part of the paper (sections 5, 6, and 7) will provide necessary elements for such a ground by identifying some central behavioral peculiarities of practical knowledge that must be explained by any empirically-adequate theory of knowledge-how. As will be seen, this is, above all, its *context-bound*, *impenetrable*, and *implicit* nature, as well as the *automatic* and *continuous* forms of processing that are connected to it. These five peculiarities will, in turn, be illustrated by examples stemming from the realms of sensorimotor knowledge (Milner/Goodale), intuitive knowledge (Damasio), and expert versus novice knowledge (Anderson), among others. We proceed by proposing a possible realization for the explanatory tasks identified in the meta-theoretical part: here we will argue that it is not by recourse to (non-)propositionality in any of its different senses that the peculiarities of practical knowledge can be explained; instead, we shall argue, *conceptuality* is a more suitable criterion for demarcating practical from theoretical knowledge, and for explaining their respective peculiarities. By “explaining” the peculiarities of practical versus theoretical knowledge we do not mean a kind of logical “derivation”. “Explaining” here is rather to be understood as showing how the realization of necessary conditions for the possession of concepts coincides with those conditions that have to be fulfilled in order to achieve the step from practical to theoretical knowledge, each characterized by their respective peculiarities. In other words, we search for “*how-possible-explanations*” of the peculiarities of practical versus theoretical knowledge. The driving role of conceptuality would also explain, in that sense, why the contents of practical knowledge cannot be easily verbally expressed, let alone abstractly represented. Such abilities only enter the scene, we argue, when knowledge reaches the conceptual level.

¹ Ryle, in his seminal approach, uses the term “knowing how” instead of “knowledge-how”. We don’t follow his usage because we think, contrary to Ryle, that know-how-phrases ascribe *genuine* knowledge, i.e., knowledge of truths (see section 2).

2 The shortcomings of intellectualist approaches

Ryle (1949), in his seminal work on knowledge-how, established a tradition of thinking that knowledge-how, as opposed to knowledge-that, is essentially characterized by its *non-propositionality*. That an action is intelligent, and thus embodies practical knowledge, comes not in virtue of its being “controlled by one’s apprehension of truths”, according to Ryle, but instead in virtue of its manifesting an ability or a disposition. Thus, Ryle’s notion of propositionality of knowledge is from the start coupled with a specific model of knowledge-application. Since this model cannot be true, practical knowledge cannot be employed by applying propositions. Indeed, if a person, in order to apply knowledge had first to “consider a proposition”, stored in his or her memory, this very act of considering a proposition would itself be an instance of practical knowledge and thus would be in need of a further act of considering a further proposition, and so on ad infinitum. Note that this means, at most, that practical knowledge cannot be manifested by virtue of *this* sort of application of propositions. But, as Fodor has remarked, “[if] the intellectualist says that, in tying one’s shoes, one rehearses shoe-tying instructions to oneself, then the intellectualist is wrong on a point of fact” (1968, p. 631). Thus, in order to avoid the whole debate turning out as a non-starter, we first have to disentangle the claim of propositionality of practical knowledge from the Rylean model of knowledge-application. But in what other sense, then, could practical knowledge be propositional?

The answer is that practical knowledge could be propositional in the sense that a person has practical knowledge by virtue of there being a rule that has a symbolic, language-like (“propositional”) representation, which is not accessible to consciousness, and which is not in need of being consciously “considered” in order to be applied in action. The knowledge embodied by this rule is instead applied in action by means of some kind of sub-personal processing of the representation. Fodor (1968) has defended such an intellectualist answer to Ryle’s

challenge by suggesting that the non-conscious representation governing the application of practical knowledge embodies “tacit knowledge”; since such tacit knowledge is applied by means of automatic mechanisms (not by intentional acts), it cannot fall victim to Ryle’s regress argument.

If we ignore the vagueness of this reading with respect to the *units of processing* in which this symbolic representation should appear, the foregoing may be a good answer to the question of how practical knowledge could possibly be propositional knowledge. In the eyes of Stanley (2011b), a more general conclusion could be drawn. According to him, since this argument that knowledge-representations need some automatic mechanisms (and not something like “considering” a proposition) in order to be applied in action, is true irrespective of the kind of knowledge involved, symbolically represented or not, all kinds of knowledge are completely *on a par* with respect to their representations—whatever they are—having to play some functional roles, mediated by an automatic mechanism, in order to be applied in action. Thus, Ryle’s analysis, according to which practical knowledge has a *dispositional* nature, can be accepted, but only at the price of accepting it for all sorts of knowledge. As such, not only can practical knowledge be propositional, but the whole distinction between propositional and non-propositional knowledge turns out to be irrelevant for characterizing sorts of knowledge, and *a fortiori* cannot be used to ground the distinction between practical and theoretical knowledge.

In other words, it is important to hold apart the thesis that knowledge is propositional in the sense of its being based on language-like representations, accessible to consciousness or not, from the empirically implausible Rylean model of knowledge application, which presupposes an act of “considering” a proposition. If we keep this distinction in mind, we find that propositionality *per se* does not provide a criterion for the theoretical versus practical knowledge distinction. Instead, all kinds of knowledge have to be “dispositional” in some sense, irrespective of their being based on symbolic, language-like representations or not.

Some anti-intellectualists, following Ryle, use the notion of “propositionality” of knowledge to refer to the fact that a person has *conscious access to linguistic propositional representations* (that is, that a person has sentences “in her mind”). Thus, for example, Michael Devitt, in a recent paper (Devitt 2011), argues that intuitively “to attribute any propositional attitudes to the ant [who has the skill of finding its way back to its nest by virtue of some neural processing] simply on the strength of that competence seems like soft-minded anthropomorphism” (Devitt 2011, p. 208). But the impression of anthropomorphism only occurs if we constrain the notion of a propositional attitude to refer to a conscious act by which a person relates to a linguistic propositional representation. The impression disappears as soon as we replace this interpretation of “propositional attitude” with a version in which the “proposition” is a rule, represented by symbolic encoding to which the ant is related by virtue of her neural mechanisms processing this encoding (or by virtue of her neural mechanisms being structured in such a way that they realize some implicit rule). That the ant “grasps a proposition” appears to be a strange description only under the presupposition that guidance by propositions implies the conscious possession of linguistic entities.

Moving from these “intuitive” considerations to arguments from the “science of knowledge-how” (cf. Devitt 2011, p. 207), Devitt identifies a “folk distinction between knowledge-that and knowledge-how” with the “psychological one between ‘declarative’ and ‘procedural’ knowledge” (2011, pp. 208-209). Now, declarative knowledge, according to Devitt, is characterized (according to what he sees as a consensus in psychology) by *conscious representation* of what is known (cf. Devitt 2011, p. 210). For example, a person has declarative knowledge of arithmetic rules only if she consciously represents those rules. Concerning procedural knowledge, Devitt refers to the distinction from computer science between “processing rules that govern by being represented and applied and those that govern by being simply embodied, without being represented” (2011, p. 210). Since

there is, according to Devitt, no decisive empirical evidence to tell us whether skills involve representations of the governing rules or not, he takes the recent picture that psychology paints of procedural knowledge “as constituted somehow or other by embodied, probably unrepresented rules that are inaccessible to consciousness” (Devitt 2011, p. 213). Finally, he argues that empirical evidence from cognitive ethology confirms this distinction between declarative and procedural knowledge by indicating that the “surprisingly rich cognitive lives” of desert ants, western scrub jays, or bottle-nosed dolphins can be understood as based on forms of procedural knowledge (to be identified with the folk notion of “knowledge-how”), but not on declarative knowledge (“knowledge-that”).

Thus, surprisingly, the anti-intellectualist Devitt and the intellectualist Fodor would agree to subsuming sub-personal knowledge, whether represented in explicit or implicit form, under the heading of knowledge-how. But the first would classify it as non-propositional, the latter as propositional knowledge. The real dissent seems to be about the question whether representations being *conscious* (and being accessible in linguistic form) or *non-conscious* makes a relevant difference for sorts of knowledge. We think that conscious availability/unavailability expresses a relevant difference for sorts of knowledge, but a difference *that can only be explained* by recourse to some fundamental distinction between practical and theoretical knowledge. Phenomena indeed indicate that the boundary between practical and theoretical knowledge coincides pretty well with conscious availability/non-availability. But Devitt’s distinction just *repeats* this phenomenon, rather than explaining it.² What we look for is a deeper reaching distinction that would be able to explain phenomenal differences such as conscious availability/non-availability and, as a consequence, verbalizability/non-verbalizability.

2 In the same way, Adams (2009) argues for a knowledge-that/knowledge-how distinction on the grounds of empirical evidence that takes recourse to experimental findings showing that declarative and procedural memory can operate independently from each other. We think that such empirical phenomena constitute explananda of the searched-for distinction, but cannot provide decisive evidence for the existence of a fundamental difference between knowledge-how and knowledge-that.

Thus, the propositionality criterion appears again unsuited for drawing an empirically-interesting distinction between practical and theoretical knowledge. As far as the intended distinction concerns the *transfer of knowledge into action* (this aspect is exactly that to which Ryle's distinction refers), *ways of representing knowledge* seem to be "on a par" and thus insensitive with respect to the distinction.

According to Stanley, it is the *semantic* notion of propositionality, with respect to which all sorts of knowledge can be subsumed as "propositional" (knowledge-that), be they based on conscious or non-conscious representations, by explicitly represented or simply embodied rules. Thus, Stanley has argued that the way in which a piece of knowledge is *implemented* (or *represented*) has nothing to do with a distinction between two *kinds* of knowledge. Therefore, the distinction between "declarative" and "procedural" knowledge as it is widely used in psychology should not be misunderstood, according to him, as providing some ground for the knowledge-that versus knowledge-how distinction: "the latter is a putative distinction between two *kinds of state*, rather than a distinction between *two ways of implementing a state*" (cf. Stanley 2011b, p. 151). Paradigmatic examples of practical knowledge, in the sense of knowledge being manifested by intelligent conduct, could turn out to be represented in a language-like way (without any conscious mediating act of "considering a proposition"), whereas clear examples of theoretical knowledge could fail to have any language-like representational background.

Stanley's *semantic* reading of propositionality is concerned with the reference of "know how"-phrases by which we ascribe knowledge-how to persons. According to our best available linguistic theories, as Stanley argues, know how-phrases have to be understood to refer to propositions. But this fact, in the first instance, does not include anything about the role those propositions play in the intelligent action of a person who knows the propositions. In particular, it does not follow that such a person possesses language-like symbolic representations that guide the person's intelligent action, or that such a person "considers" the proposition

in order to apply his knowledge in action. If the correct understanding of the semantics of knowledge-phrases, no matter whether it is theoretical or practical knowledge that is ascribed by them, is that they refer to propositions, then this propositional nature of knowledge, according to this reading, cannot be used to draw any distinction between theoretical and practical knowledge.

Now, someone could object that Ryle's distinction is concerned with the *nature* of knowledge, e.g., how knowledge is represented in a person, but not with what is involved in knowledge *ascriptions*. Thus the semantic reading of propositionality would be irrelevant for the theoretical versus practical knowledge distinction. But note that Ryle's analysis of practical knowledge actually starts by asking questions like: "When the person is described by one or other of the intelligence-epithets" (Ryle 1949, p. 28), what sort of knowledge is this description imputing to the person? That is, Ryle asks for the semantics of knowledge-ascriptions for typical cases of practical knowledge. Therefore, it is not at all clear that a semantic reading of propositionality is irrelevant for his analysis. On the contrary, the sense in which Ryle is concerned with the "nature" of knowledge is expressed, by him, by means of an analysis of the role that knowledge-phrases play in actual linguistic practice.³

It has now been shown that both possible readings of "propositionality", that is, the *representational* and *semantic* readings, are *relevant* for Ryle's proposed theoretical versus practical knowledge distinction, but neither is suited to grounding the distinction: Whether a piece of knowledge is a case of practical or of theoretical knowledge does not depend on whether it is supported by language-like structures or not; and, since *all* knowledge is semantically propositional (if Stanley is right) it does not depend on its semantic propositionality either.

³ Contrary to this, Noë (2005) argues that "Ryle's distinction is not a thesis about the sentences used to attribute propositional and practical knowledge, respectively". He claims that "Ryle was not an ordinary language philosopher". How then, would Noë, for example, understand Ryle's appeal to linguistic use in his deflationary account of the "will"?

Thus, it seems as if no criterion for the distinction between practical and theoretical knowledge could be available from the intellectualist point of view. But, we shall see that, from Stanley's revised dispositional analysis of knowledge, rather surprisingly a new possible criterion emerges. Let us, therefore, follow the path of this analysis, which is intended by the author to show how, contrary to Ryle, the (semantic) propositional nature of knowledge is compatible with its dispositional nature.

According to Stanley (2011b), even if we accept Ryle's general claim that knowledge has to be understood as dispositional,⁴ "there still need to be automatic mechanisms that mediate between dispositions (and abilities) and the manifestation or execution of these dispositions and abilities" (Stanley 2011b, p. 26). What has to be true of theoretical knowledge, namely the existence of mechanisms that mediate the application of that knowledge, has to be also true of practical knowledge. The complex of dispositions on which the ability to catch the fly ball rests may be completely intact, even if the player sometimes does not succeed in catching the ball because he has become tired or has momentarily lost concentration. When that happens, his executing mechanisms can fail. As has often been identified in the debate on knowledge-how, having the right dispositions (and thus having the right sort of practical knowledge) does not always guarantee successful performance (cf. Snowdon 2004).

Even if, from the intellectualist point of view, all forms of knowledge—be they "practical" forms of knowledge or not—could be, and indeed have to be, analyzed with respect to their dispositional nature, the question seems, by the very phenomenology of practical know-

ledge, to be more urgent than for cases of theoretical knowledge: How can knowledge be dispositional and propositional at the same time? Stanley & Williamson (2001) have suggested that cases of practical knowledge can be captured by means of a "practical mode of thinking", by which a person who has practical knowledge has access to propositional contents. If, for example, a person knows *that a certain way of riding a bike is a way for her to ride a bike*, then her thinking of that proposition is in a peculiar way self-directed, it is a "first-person-way" of thinking the proposition. Stanley (2011b) has developed this suggestion further into a dispositional theory of knowing a proposition.

Gareth Evans (1982), in his analysis of "demonstrative knowledge", has provided a useful framework of first-person dispositions: My thinking is a demonstrative belief about a perceptually-presented object if I will be disposed to have changes in that object affect my belief (Stanley 2011b, p. 110). Thus, my thinking of an object in the world as "myself" involves a permanent disposition to let my thoughts and actions be determined by my own bodily perceptions. Now this schema seems to fit the practical way of thinking that occurs when it comes to propositions like "This way of riding a bike is a way to ride a bike for me": A person manifests knowledge of this proposition by, while riding a bike, manifesting the disposition to react to certain kinesthetic sensations mediated by her own bodily movements by means of adequate motor commands.

We accept this as an adequate way of describing the phenomenological peculiarity of "practical ways of thinking" a proposition. Indeed, when described in this way, practical knowledge can be propositional and dispositional at the same time. But this analysis does not tell us—and indeed is not meant to tell us—how the distinction between practical and theoretical knowledge can be grounded. That there is such a distinction seems obvious *inter alia* on the basis of the functional characteristics peculiar to practical knowledge, such as its domain-specific nature, its limited transferability, its non-penetrability, and so on. Stanley's

⁴ Contrary to what Noë (2005) has claimed, Stanley thus does not attack Ryle's identification of "knowledge how" with the possession of abilities *tout court*. What Stanley objects to is the supposed opposition between knowledge as the possession of abilities and propositional knowledge on which Noë, assuming that propositional knowledge necessarily entails understanding of propositions, insists. Even the earlier work (Stanley & Williamson 2001) tries to account for the dispositional nature of practical knowledge by introducing the concept of a "practical mode of thinking". On the contrary, any unrestricted identification of knowledge-how with abilities confronts the problem of how to account for cases in which practical knowledge survives the loss of ability. The distinction between dispositions and their manifestation by means of executing mechanisms accounts for this problem.

dispositional theory fails, at least at first sight, to deliver any resources for *explaining why* practical knowledge is distinct from theoretical knowledge on the basis of these functional characteristics. The main shortcoming of recent intellectualist approaches, in our opinion, is not that they simply neglect the peculiarities of practical knowledge. Rather they are deficient insofar as they do not provide an explicit positive demarcation criterion of practical versus theoretical knowledge that would go beyond capturing the well-known phenomenological peculiarities and make it compatible with the proposed fact that all knowledge is (semantically) propositional. Before we go back to Stanley's analysis, in order to show how some explicit demarcation criterion could possibly be drawn from it, we ask whether recent anti-intellectualist approaches do a better job of providing a demarcation criterion.

3 The shortcomings of anti-intellectualist approaches

The anti-intellectualist position has recently been supported by, among others, [Toribio \(2008\)](#), [Young \(2011\)](#) and [Newen & Jung \(2011\)](#). Newen and Jung assume that Ryle's distinction between knowledge-how and knowledge-that should be taken as referring to the *nature* of knowledge. From a naturalist point of view, the most general way to characterize knowledge is to say that it is based on mental representations. Thus the distinction between practical and theoretical knowledge, from that perspective, has to be spelled out as a distinction between *ways of representing* something, or between representational formats. Now, theoretical knowledge, according to [Newen & Jung \(2011\)](#), can be identified with the propositional representational format, whereas they hold that practical knowledge comes in two (non-propositional) varieties, one characterized by the format of *sensorimotor* representations, and the other by what they call *image-like* representations.

Concerning the first of these representational formats, namely the propositional format, we are confronted with the same problem we

faced when considering Ryle's notion of propositionality. What does it mean to say that a representation is propositional? It should not mean that the content of the knowledge is or can be made available to the person in the form of consciously-accessible *linguistic structures*. Even if the property of *explicitness vs. implicitness* of knowledge is often used to distinguish between theoretical and practical knowledge, this merely descriptive criterion does not help to *explain* the theoretical versus practical distinction, but preferably should be explained by the more principled criterion we are looking for. If, on the other hand, we take "propositional" to mean that the kind of *processing* connected to a piece of knowledge has a language-like structure, how do we identify the units of processing to which this characterization is supposed to refer? Even if it were possible to precisely identify the level of processing that accounts for propositionality, it would be far from clear how the characteristics of theoretical versus practical knowledge could be explained by means of that supposed representational fact. As we have already pointed out in discussing Ryle's notion of propositionality: Why should it be the case that "theoretical" knowledge is necessarily connected to propositional representations, and, correspondingly, practical knowledge to non-propositional ones?

According to [Young \(2011\)](#), what we call "knowledge-how" may appear in different forms, which are accompanied by more or less comprehensive linguistic mastery of propositions. The sort of knowledge a guitar player manifests in his playing may be either such that he is able to *articulate* that, for example, G should be played rather than G#, or such that he may only be able to *experience* his performance as appropriate guitar playing ([Young 2011](#), pp. 57f.). In the latter case, his knowing how to play guitar is constituted by specific dispositions to react in particular ways to the conscious auditory and motor experience of his own playing. Even this form of knowledge may be reducible to propositional knowledge, however, since the player is potentially able to instruct himself with the help of demonstrative pronouns denoting parts of his actual auditory experiences. Whereas

those forms of knowledge-how may, according to Young, be reducible to propositional knowledge, he thinks that there is a clear case of *irreducible* knowledge-how that is constituted by “purely” sensorimotor abilities, and that is exercised without being supported by any kind of propositional knowledge. Such kinds of sensorimotor abilities are exemplified, according to Young, by the case of DF in the Milner/Goodale-experiments.

Patient DF is impaired in her ability to recognize objects, despite showing intact basic visual processing abilities. Milner and coworkers presented to DF a letterbox in which the slot through which one inserts letters could be rotated to vertical, diagonal, or horizontal orientations. DF had problems when she was asked to visually match the orientation of the slot to different alternatives. However, when asked to actually insert a letter, she was able to reach towards the slot while orienting her hand in accordance with the spatial orientation of the slot. Thus, DF has the ability to use visual information in purposeful object manipulations without being able to consciously visually process or experience them. On the other hand, another patient, IG, showed conscious visual awareness of objects without being able to practically manipulate them. Apparently, then, there are two independent neural pathways for processing visual information: the ventral path, leading to visual identification and corresponding to conscious experience, and the dorsal path, used for non-conscious action control and execution. In pathological cases, one or the other (DF vs. IG) of these pathways does not work, whereas the other remains intact (Milner & Goodale 1995, 2008).

What is the reason for Young’s assuming that the case of DF exhibits “irreducible” knowledge-how? The reason seems to be that DF is not able to use linguistic propositions—in whatever rudimentary form—to refer to aspects of the visual scene. She simply has no conscious access to the visual scene whatsoever. Young thus takes “propositionality” of knowledge to be constituted by conscious access to—possibly rudimentary forms of—linguistic propositions. But, as we already have seen, lacking conscious access to linguistic propositions accompanying

the performance of knowledge-how does not exclude the “propositionality” of that knowledge-how in the semantic sense of “propositionality”, and neither does it exclude the “propositionality” of that knowledge in the sense of being based upon symbolic language-like processing.

Toribio (2008) gives a similar argument against the possible propositionality of DF’s knowledge. She argues that

DF has no conscious awareness of this visual information [the information available on the dorsal route] and has no phenomenal experience as to the appropriateness of her own performance, but she has proprioceptive awareness of the features that govern her visually guided action in this particular task. (cf. Toribio 2008, p. 13)

This situation, according to Toribio, is relevantly different from the example of Hannah’s knowing how to ride a bike. In the latter case, Hannah has not only proprioceptive, but also *conscious* awareness of the sensory information available. Why does this difference matter? It matters, Toribio suggests, because in order to make plausible that a person’s knowledge-how is somehow “guided” by a proposition, this guidance has to be spelled out by a real process of “entertaining” or “contemplating” the proposition by the person. Suggesting a propositional reading of Hannah and DF’s knowledge without being able to point out some possible realization of “entertaining a proposition” in these cases “threatens to make us lose our grip on what propositional knowledge is” (cf. Toribio 2008, p. 13). But Stanley & Williamson (2001), Toribio claims, are unable to provide such a possible realization in the case of DF:

DF couldn’t possibly entertain such a proposition because she cannot grasp one of its constituents – she cannot perceive the features, e.g. the orientation, that governs her motor behavior in the posting task, and hence couldn’t recognize them as in any way constituting a reason for her action. (cf. Toribio 2008, p. 9)

We think that Stanley's elaboration on "practical ways of thinking a proposition" is able to overcome this objection. We can very well understand what it means that a person thinks a proposition p without being able to sensually identify the objects constituting p . Sensual identification ("grasping") is a precondition for *conceptual* apprehension of the constituents of a proposition, but it is not a precondition for non-conceptual attitudes to propositional contents, by way of proprioceptive information.

What the performance of DF in the Milner/Goodale-experiments indeed shows is that sensorimotor processing of visual information is sufficient for entertaining practical abilities and does not require any conscious processing, in particular no *linguistic* processing, if we suppose that linguistic processing is necessarily conscious.⁵ This result does not imply that sensorimotor processing is independent of (and opposed to) propositional processing. Sensorimotor processing could use "propositional" representations, only if these propositional representations were not linguistic representations (cf. Fodor 1968). Thus, the case of DF cannot be understood as supporting the sensorimotor-propositional processing-classification of knowledge. There is still no indication that there are two independent types of cognitive processing, a propositional and a sensorimotor one, to say nothing about the possible explanatory virtues of such a distinction.

That the sensorimotor vs. propositional classification is lacking any theoretical foundation that could determine whether this distinc-

tion is already complete or has to be completed in certain ways becomes obvious when further classificatory distinctions are proposed. For example, Newen & Jung (2011) introduce, in addition to the sensorimotor and propositional format, a third representational format, called *image-based knowledge*, which they think can supplement the knowledge-how variety. An example of image-based knowledge, according to the authors, is a high jumper's generation of a mental image of his planned jump before his running up. The authors argue that the mental image can take the role of controlling the performance of the action. The action, in cases of image-like knowledge, is thus "guided" by an image, just as motor reactions to bodily experience supposedly guide actions in the case of sensorimotor knowledge, and propositions supposedly guide actions in the case of propositional knowledge. Now, we think that it is far from clear how mental imagery or real images can "guide" actions. Even if we could clarify what "guiding" here means, there is at least a possible alternative interpretation of the role of mental images in acting, namely a common cause interpretation, according to which the performance of the action and the occurring of a corresponding mental image have a common cause, namely the neural processing that is the real cause of the different aspects of the performance, which thus "guides" the action. If such an interpretation was correct, the mental image would not be a "guide", but would merely be an epiphenomenon of the processing that produces the action (cf. Pylyshyn 1984). That this alternative interpretation exists shows that there is no clear indication that "image-based knowledge" is an independent third kind of knowledge that would legitimately supplement the classification.

On the other hand, research in psychology and cognitive neuroscience indicates that it is possible for non-conscious and non-linguistic types of knowledge (e.g., intuitive knowledge) to guide actual behavior, and which cannot be classified as "sensorimotor" knowledge.⁶ As long

⁵ Note that this does not necessarily mean that there is also no *conceptual* processing involved. As Stanley points out, declarative knowledge is sometimes defined as "knowledge that can be consciously and intentionally recollected", as opposed to procedural knowledge, which is taken to be "knowledge expressed through experience-induced changes in performance" (Stanley 2011b, p. 154). This reading of the procedural-declarative distinction proposes to fix it by translating it into the "explicit" versus "implicit"-distinction, where it seems to exactly match the distinction of two pathways of processing that are exhibited in the Milner/Goodale-experiments. But it cannot be taken as grounding the theoretical versus practical knowledge distinction. We agree with Stanley, who claims that practical knowledge can have a propositional content that is able to be verbalized—the subject can be able to linguistically express what she knows. Stanley's example is that of "physicians skilled at a procedure, who are also very good at describing to others how they do it"—they "possess explicit procedural knowledge" (2011, p. 159). Thus knowledge may be procedural in the sense of the above definition, and at the same time conscious and linguistically expressible.

⁶ A further type of practical knowledge that fulfills this criterion seems to be expert knowledge in areas that are not reducible to sensorimotor processing: e.g., chess or financial stock markets.

as there is no theoretical principle or framework from which the classification of possible forms of knowledge-how can be derived, there is in our opinion no reason to exclude such types of knowledge from the knowledge-how variety.

To give an example of non-sensorimotor knowledge-how: Bechara et al. (1997) examine the behavioral, subjective, and physiological states involved in intuitive decisions. Participants played a card game (known as the *Iowa gambling task*) in which they had to repeatedly (up to 100 times) pick cards from four different decks that could lead to wins as well as losses. In the long run, drawing from some decks led to smaller or larger winnings, and others to smaller or larger losses. The goal was to maximize one's play money on the basis of a \$2000 starting sum. Unknown to the participants, the card decks were pre-organized so that all decks would lead to wins in the first few draws. During the game two good decks turned out to be relatively safe (i.e., small wins and losses) leading to an overall net win, while two bad decks turned out to be relatively risky (i.e., large wins, but also large losses) leading to overall net loss.

The hidden win-loss dynamics and relations between the outcomes allowed the researchers to separate different periods of card-drawing behavior (standing for different knowledge states) during the game. A first *pre-punishment period* stood for the phase of early wins, a second *pre-hunch period* for the phase in which subjects started to get a feeling that there were differences between decks in terms of safety vs. risk-taking, a third *hunch period* for a phase in which subjects started liking or disliking certain decks without exactly knowing why this was the case, and a last *conceptual period* in which subjects were able to articulate their preferences and the reasons for these preferences between different decks. Not all participants reached the *hunch* or the *conceptual* period of the game.

Of foremost interest were observations made in the *pre-hunch period*. Normal participants, as opposed to participants with prefrontal damage, began to develop behavioral preferences for the good and less riskier card decks during this phase, and also showed anti-

cipatory skin conductance responses (reflecting minimal perspiratory reactions standing for fear responses) when planning to draw from riskier decks, although they were not consciously aware of these preferences, or of any physiological reactions during this phase of the game. Showing these non-conscious and involuntary responses during the *pre-hunch period* was prerequisite for subjects to advance to the *hunch* as well as the later *conceptual period*. A control group of prefrontally-damaged participants⁷ did not show any of the described physiological skin responses during the experiment, and their card-drawing behavior as well as their subjective reports showed no sign that they had developed knowledge of the riskier behavior associated with picking cards from certain decks.

The intuitive knowledge that is reflected in this study makes up for a further possible form of knowledge-how (for other examples of intuitive knowledge see Myers 2004; for intuitive core knowledge about geometry, numerosity, and ordering see Spelke 2000; for intuitive knowledge of experts see Dreyfus & Dreyfus 1986). Instead of adding new forms of knowledge-how in some arbitrary way, we think that it is more promising to look for a general criterion for knowledge-how that has the potential to explain the salient characteristics of knowledge-how, and at the same time is suited to give a framework for the possible surface forms in which knowledge-how may appear, including the sensorimotor and intuitive forms described above. We suppose that the most promising candidate for such a criterion is *non-conceptuality*.

4 How can propositional knowledge be non-conceptual?

How can it be true that the knowledge held by a person is “propositional” in its semantic sense⁸ without being conceptual? Would not the per-

⁷ Several studies (e.g., Barch et al. 2001; Bechara et al. 2000; Halligan et al. 2004; Stuss & Alexander 2007) indicate that lesions of the prefrontal cortex can lead to a number of cognitive and affective problems, most notably working memory problems, deficits in executive functioning such as planning, goal selection, task monitoring, deficits in inhibiting thought and action impulses, problems in outcome anticipation, and risk-taking behavior.

⁸ Note that we have accepted Stanley's thesis that all knowledge is propositional in *that* sense.

son necessarily need a grasp of the concepts a proposition is “composed of” in order to have knowledge of that proposition? The answer is that in order to have *conscious* knowledge of a proposition given in *linguistic* form it is necessary to have a grasp of the concepts of which the linguistically-given proposition is composed. But Stanley’s notion of knowing a proposition is not restricted to linguistically-given propositions. For example, if Hannah knows the proposition that “this way is a suitable way for me to ride a bicycle”, her way of knowing this proposition is a *practical* way of knowing that does not include knowledge of linguistic entities, but shows up by manifesting dispositions to react to certain kinds of bodily experiences. Thus, as much as knowledge-how is involved, it is possible to have knowledge of a proposition without being able to grasp the concepts the proposition is “composed of” when given in a linguistic form. The case can be made plausible by looking again at the Milner/Goodale-experiments: although the patient DF knows “how to put a card into a vertical slot”—and thus knows a proposition—due to a defect in her ventral pathway she is not able to have a conceptual understanding of the linguistic components of that proposition.

Stanley (2011b) has formulated objections to conceptions of non-conceptual content, at least when they are directed against propositionality *tout court*, as for example in Dreyfus (2007), according to whom “embodied skills [...] have a kind of content which is non-conceptual, non-propositional, non-rational [...]” (p. 360). His main argument is that ascriptions of knowing-how create opaque contexts (Stanley 2011b, p. 168). But this argument does not seem very strong, if seen from Stanley’s own perspective of a dispositional reading of ways of knowing a proposition in the case of knowing-how. How the objects occurring in the propositional content are conceptualized does not make any difference to the subject’s knowing the proposition, namely his being disposed to react to his own bodily experiences in a certain way (think of the guitar player). Thus, the dispositional reading of propositional knowledge is simply not compatible with the proposed fact that proposi-

tional contents are individuated by concepts. Instead, it implies that, in case of knowledge-how, persons have propositional knowledge that is indeterminate with respect to any conceptualization of the objects occurring in the propositional content. We therefore object to Stanley’s claim that “I cannot be said to know how to ride a bicycle if I have no clue what a bicycle is” (Stanley 2011b, p. 170). Someone can be able to manifest a well-determined disposition with respect to riding a bicycle, whatever conceptual understanding, if any, he has about bicycles.

In face of the DF-case in the Milner/Goodale-experiments, Stanley admits that:

[...] DF cannot accurately report on the orientation of the slot, whereas the normal agent can. DF’s knowledge of how to put a card into a slot is propositional knowledge that is based on a non-conceptual understanding of the orientation of the slot, understood here in the sense of an understanding of the orientation of the slot that is not available to conscious apprehension. She is able to have propositional attitudes about a way of posting a card into a slot in virtue of this non-conceptual understanding of orientation, yielded by her intact dorsal processing pathway. In contrast, the normal agent does have consciously available knowledge of the orientation of the slot before she acts. This is a difference between DF and the normal agent, but not one that can be used to deny that DF’s action is guided by propositional knowledge of how to put a card into a slot. (Stanley 2011b, p. 172)

In the remaining sections, we will follow the path opened by the suggestion that knowledge can be propositional without being conceptual. Whereas we hope to have shown that the propositional/non-propositional-distinction is not fruitful for explaining practical knowledge, we argue that the conceptual/non-conceptual distinction does have this potential. The idea, following Stanley’s proposal, is that knowledge-

how is, in general, knowledge of propositions by way of non-conceptual understanding. But we do not stick to the definition of “conscious apprehension” that in the DF-case indeed coincides with conceptual grasp. There can be conceptual grasp even in the absence of conscious apprehension (as it seems to be the case for certain animal species where the presence of consciousness is at least doubtful). Instead we take recourse to a minimal conception of “conceptuality” that has been developed by Newen & Bartels (2007) in the context of animal concepts. This minimal conception does not depend on consciousness. First, however, we shall explore the already-noted peculiarities of practical knowledge. It is these peculiarities that a fruitful conception of knowledge-how, based on the contrast between “conceptuality” and “non-conceptuality” needs to be able to explain.

5 The peculiarities of practical knowledge

An adequate meta-theory of human knowledge should be able to account for empirical differences observed when people use practical rather than theoretical knowledge in the most general terms, and be able to deliver an explanation for these differences. The starting point for the need to distinguish between practical and theoretical knowledge is the behavioral and neurological differences or dissociations in performance in different sensory, motor, or cognitive tasks, e.g., performance differences between experts and novices, between normal and prefrontal patients, between DF and IG. In actual research observed behavioral or neurological differences and dissociations are often accounted for by describing them in terms of polar opposite knowledge attributes or effects. In our understanding this is a first step in the direction of a theory of knowledge-how, even if it is still short of delivering a satisfactory explanation of the observed behavioral and neurological differences.

In the cognitive science and psychological literature, one finds the following polar opposite ascriptions of attributes of knowledge-how as opposed to knowledge-that:

A. Context-bound versus context-free knowledge. Knowledge-how is specific to the domain or the situation of its use, whereas knowledge-that is not. In other words, knowledge-how is about *situational* skills, while knowledge-that is about *general* facts (e.g., Clark 1997; Clancey 1997). For example, throwing a javelin and anticipating its movement when it leaves the hand is a case of context-bound knowledge, whereas calculating the biomechanical forces needed for optimal performances (e.g., the ballistics of an optimal flight trajectory) is an instance of context-free knowledge. Chess experts as compared to novices have superior context-bound knowledge of constellations of chess figures, which helps them to reproduce specific shortly-presented board situations from memory. However, their superior knowledge does not help expert chess players to reproduce random constellations of chess figures from memory, as their skill for applying context-bound perceptual chunking mechanisms on meaningful constellations of figures does not prove beneficial.

B. Impenetrability versus penetrability of knowledge. Knowledge-that is penetrable by other cognitive processes or meta-processing, whereas knowledge-how is impenetrable (Pylyshyn 1984, 1990). Impenetrability means that use of knowledge-how is not changed by internally (e.g., beliefs, goals) or externally (e.g., distracting stimuli) triggered cognitive processes. One example is subitizing, i.e., the rapid, accurate, and confident estimation of the number of displayed elements (e.g., stones), which works fine and is robust against internal or external distractions. In contrast, the use of knowledge-that to determine the number of regularly arranged objects by counting them or doing mental arithmetic (e.g., adding over rows of elements $3+5+4+2+\dots$) is prone to interferences from internally- or externally-activated cognitive processes. If athletes change the order of different sensorimotor sub-processes (e.g., in technical sport disciplines such as high-jumping or hitting a golf ball), they can encounter considerable problems and might need additional time and effort to build up new

knowledge-how. Not so well-trained movements (e.g., dancing steps in beginners) can be more easily rearranged.

C. Implicit versus explicit knowledge. Use of knowledge-how takes place largely outside of awareness and hence cannot be verbalized, while knowledge-that is to a large degree consciously available and can be verbalized. In the last decades psychological research has made substantial progress in distinguishing between implicit and explicit forms of human learning, memory, and information processing (e.g., [Dijksterhuis & Nordgren 2006](#)). People learn the grammar of natural language or internalize their society's norms implicitly, that is, without conscious knowledge of the principles that guide their language use or their social behavior (e.g., [Reber 1989](#)). Implicit memory is, for example, displayed in cases of amnesia, in which patients are not able to explicitly recall previously-presented items or events from memory, while performances on tasks that do not require explicit memory such as perceptual priming or sensorimotor skills are undisturbed and virtually normal (e.g., [Tulving & Schacter 1990](#)).

D. Automatic versus effortful processing. Use of knowledge-how is automatic in the sense that it requires little attentional monitoring or guidance, and in the sense that its demands on working memory are quite low ([Bargh & Chartrand 1999](#)). Use of knowledge-that is generally more effortful, and can be shown to require significant attentional as well as working memory resources ([Hasher & Zacks 1979](#)). Good examples of the distinction between the automatic and effortful use of knowledge can be found in the domain of spatial cognition: Blindfolded navigators (animals as well as humans) complete triangles by returning to the starting point on the basis of automatic vestibular and kinesthetic path-integration mechanisms (knowledge-how), while only humans are able to use effortful geometrical calculations (knowledge-that) to find their way back to the origin of the outbound travel. Experiments show that simultaneous secondary tasks (e.g., to-be-ignored spatial movements vs. counting operations) differentially affect the one or the other type of knowledge processing ([May & Klatzky 2000](#)).

To give another example from research on spatial cognition: Wayfinding on the basis of multimodal sensory inputs from the surroundings and from automatic updating is very different from the quite effortful and highly disturbable use of knowledge-that that results from listening to verbal route-descriptions or maps ([Montello 2005](#)).

E. Continuous versus discontinuous processing. Use of knowledge-how expresses itself in smooth and continuous processing, while knowledge-that is normally reflected in step-by-step processing along a discontinuous path of intermediate knowledge states. Recent dynamic systems accounts of the sensory, motor, and cognitive processes underlying human knowledge use describe these differences in terms of different attractor landscapes of mental or neural state spaces ([Spivey 2008](#)). Research into children's cognitive development, for example, reveals that there are two levels of spatial location coding in memory. In a first phase, children learn to code the metric distance between locations (e.g., allowing them to find previously hidden objects in terms of distance from the sides or the corners of a rectangular sandbox). In a second phase, children attain the ability to impose organization on their spatial knowledge (e.g., allowing them to divide the spatial layout in hierarchical subsections or regions). The shift from the first to the second level reveals itself in changes in the types of spatial errors (discontinuous vs. continuous distributions) children commit when locating hidden objects ([Newcombe & Huttenlocher 2000](#)).

This list of opposing attribute pairs is probably not complete, but seems a good starting point for our purposes. It can be thought of as a general description and characterization of practical knowledge in contrast to theoretical knowledge. Not every single case of knowledge use will be easily describable by means of the list, or will even require a full description along all opposing attribute pairs. However, chances are good that the overwhelming majority of cases will be adequately described by using such a set of opposing attributes, and, generally, the profile over the five attributes will correctly apply. We will ar-

gue that this list of attribute pairs, together with their predominant assignment to the one or other knowledge variety, is what an adequate and fruitful theory of knowledge-how vs. knowledge-that should be able to account for.

6 Conceptuality as a demarcation criterion for knowledge-that versus knowledge-how

We propose conceptuality as a demarcation criterion for knowledge-that in relation to knowledge-how that is able to account for the peculiarities of both knowledge types outlined in the last section. In order to show that conceptuality can do the job, we have, in a first step, to establish a notion of *concept* that does not *presuppose* in an obvious way characteristics of knowledge-that, i.e., the notion we look for should not entail that concepts are essentially linguistic entities enabling persons to verbally express knowledge-that. What we then need, in a second step, is a notion that entails some fundamental and (hopefully) non-contentious assumptions about necessary conditions for concept possession in terms of abilities. Characterizing concepts in the form of abilities necessary for concept possession should enable us to show that having those abilities necessary for concept possession is exactly what is needed for the subject to overcome the peculiar limitations accompanying knowledge-how, and thus to gain access to the level of knowledge-that (see section 5).

In shaping the sought-for notion of conceptuality we take recourse to work by Allen & Hauser (1991), Pylyshyn (1990), and Newen & Bartels (2007). Allen and Hauser have claimed that, from the perspective of interpreting the behavior of systems including human and animal organisms as much as artificial systems, the ascription of genuine concepts requires “evidence supporting the presence of a mental representation that is independent of solely perceptual information” (Allen & Hauser 1991, p. 231). The criterion of independence, as called for by these authors, is that it enables the system to show flexible

behavior, in contrast to the performance of rigid mechanisms:⁹

[I]ndependence in this sense entails that the responses of the animal to a certain stimulus are not just ‘driven by’ that stimulus, and are also not to be explained as cases of stimulus generalization, i.e., discrimination by a mechanism responsive to a single basic stimulus. (Newen & Bartels 2007, p. 287)

If the reactions of a system to a given stimulus can be modified by the presence of additional stimuli representing the peculiarities of the situation in which the reaction occurs, the system will be first able to *generalize*—as rudimentary as that ability may be—the information received. It is then that we can legitimately ascribe the possession of concepts: “First, an organism whose internal representations are concept-like should be able to generalize information obtained from a variety of perceptual inputs and use that information in a range of behavioral situations” (Newen & Bartels 2007, p. 287).

We thus arrive at a *criterion for conceptuality*, which can be called, following Allen (1999), the “transcendence of particular stimuli” or, in terms given by Pylyshyn (1990), the “criterion of informational plasticity”. Essentially the criterion requires the “possibility of the modification of a response in the light of additional information” (Allen 1999); the kind of response has to depend, crucially, on other sources of information (cf. Newen & Bartels 2007, p. 287).

The criterion considered above is still not sharp enough. As long as we do not further specify the “modification of a response” occurring “in the light of additional information”, each sort of extension of the processing capacities of an individual would count as reaching the level of “conceptuality” if only this extension enables the individual to integrate some additional

9 One example of a “rigid mechanism” is the behavior of ants responding to the presence of acidic byproducts from the decomposition of dead con-specifics: in tests they rigidly remove anything from the nest that is painted with oleic acid, even live con-specifics.

source of information into its behavioral repertoire. Thus *diversification* of processing capacities could then not be distinguished from *transition* from non-conceptual to conceptual processing capacities.

To get a criterion for conceptuality it is required that the “modification of a response” mentioned above concerns *classificatory* behavior. The “additional stimulus”, in that case, not only has to work as a switching point, opening one or the other pathway for a response *within* a non-conceptual behavioral pattern, it also has to stand for a *category*, according to which the actual behavioral pattern can be classified.

One example of additional stimuli characteristics standing for a category has been described by Newen & Bartels (2007) with respect to the conceptual abilities of the grey parrot Alex (Pepperberg 1999). In order to be able to form elementary color concepts, for example the concept “green”, Alex should not only be able to generalize over a class of similar stimuli and thus to identify a sample of different green objects, but should additionally be able to represent green *as a color*. Only then could we ascribe to him the ability to classify green objects according to a well-determined class concept.

The test items by which Pepperberg examined the classification abilities of the animal were, for instance, “What color?” or “What shape?”. These questions should

[...] determine if he [Alex] could respond not only to specific properties or patterns of stimuli [e.g., to green objects], but also to classes or categories to which these specific properties or patterns belong [...]. Could he, for example, go beyond recognizing what is, or is not, ‘green’ to recognizing the nature of the *relationship* between a green pen and a blade of grass? (cf. Pepperberg 1999, p. 52)

It happened that Alex was indeed able to classify the given “key” stimulus, e.g., a green, round object, visually presented to him, as “green” or “round” according to different dimensions (e.g., color or shape) represented by

additional auditory stimuli. His choice of response (“green” or “round”) turned out to depend crucially on the “additional information” given in form of the auditory stimulus. As such, Newen and Bartels concluded that “Alex was able to represent different properties while having only one and the same visual input of an object.”¹⁰

With this example in view, Newen & Bartels (2007) formulated the following requirements for the possession of concepts—for instance, the concept “red”: A cognitive system has the concept “red” only if (i) it has relative stimulus independence such that it depends on some additional mechanism—which detects and weighs stimuli other than the key stimulus of redness—to determine that the system focuses on redness while perceiving a red square, in contrast to some other property; and (ii) the property of being red is represented as an instance of the dimension “color”.¹¹

Note that the above-mentioned definition of conceptuality does not only require the existence of some “additional stimulus” to which the individual has to be responsive, but that there has to be some additional internal *mechanism* of processing by which the individual is able to “detect and weigh” a specific additional stimulus as standing for a *category* (e.g., “color”). The responsiveness of the individual to that stimulus shows up when it focuses its attention on those aspects of a scene, or to those items of a behavioral pattern, which exemplify the respective category.

Another example would be the balancing of coffee cups by a waiter in a restaurant. Let us assume that the waiter for some time possesses the ability to balance cups of different shapes without spilling coffee, and without consciously attending to a particular cup, or the

¹⁰ Cf. Newen & Bartels (2007), p. 293. That the auditory stimuli “What color?” or “What shape?” were really understood by Alex as asking for the respective category was tested by Pepperberg using additional auditory signals of the form “What’s same?” and “What’s different?” The correct response would be the label of the appropriate category, e.g., the mastery of categories could be verified in the sense that Alex successfully identified the essential functional role of category terms like “color” or “shape” as dividing the objects of the world into “sameness” equivalence classes.

¹¹ Cf. Newen & Bartels (2007), p. 296. These conditions are only two of a total of four conditions. But only these two matter with respect to our discussion.

shape of a particular cup that he is currently dealing with. At some point he is told that there are essentially two different kinds of cups, one high and cylindrical, and the other flat and bowl-shaped (this information is the “additional stimulus”). The waiter “detects and weighs” the additional stimulus by focusing his attention, from that time on, to his own specific handling of cups, depending on the sort of cup a particular exemplar belongs to. He might then detect that he had previously managed to deal with both kinds of cups efficiently and without spilling coffee without even noticing that liquids in both reacted in different ways to his movements. The waiter’s behavior has switched from a former “non-conceptual” dealing with coffee cups to a form of behavior that is “conceptual” in the sense of exhibiting an additional ability, namely the ability to classify his own performance in balancing coffee cups according to a category (in this case the cups’ shape).

How does such a notion of conceptuality relate to Evans’ notion of non-conceptual knowledge in terms of first-person dispositions that we made use of in sections 2 and 4? If the possession of concepts is constituted (in contrast to non-conceptual cognitive processing) by the gaining of additional abilities, it should be made plausible how those additional abilities connect to a non-conceptual basis in Evans’ theory.

In our treatment of his theory, we followed an interpretation of Evans’ work as implying that non-conceptual knowledge relies on the disposition to have one’s own motor reactions be determined by sensory and kinesthetic information that is mediated by either some external object or by one’s own body. Again, the waiter dealing with the coffee cups may help to illustrate the point. The waiter’s experienced handling relies on a disposition to have his motor actions determined by the multimodal sensory information that is mediated by holding coffee cups in his hands. The waiter’s knowledge-how to balance the cup might be completely independent of any conceptual reference to coffee cups. He could be the experienced waiter that he is—at least with respect to his balancing ability—without even knowing in a conceptual way “what a coffee cup is”. Reference to the ob-

jects he is dealing with was accomplished only by being able to react in a coordinated way to sensorimotor information originating from handling these objects.

At the time he is told that coffee cups come in two different shapes, his cognitive system enables him to use that information such that he begins to rely on a category (i.e., a cup’s shape) in order to refer to coffee cups, and to classify his own balancing behavior according to the objects thus categorized. He reaches, in some minimal way, the level of conceptual knowledge, since he now begins to identify both, that is, the objects and his behavior with respect to these objects, by conceptual means.

7 Explaining the peculiarities of knowledge-how by means of conceptuality

Equipped with an adequate notion of conceptuality, we now proceed to show that concept-possession is exactly what is needed for a cognitive system to overcome the specific limitations associated with knowledge-how, and hence be able to gain access to the level of knowledge-that. Why exactly is it necessary for a system to possess concept-like representations in order to have knowledge-that as opposed to knowledge-how?

1. *Context-bound versus context-free knowledge.* For this polar contrast the answer, in short, will be that conceptual representations are precisely those representations which make the subject able to generalize information over a range of different behavioral situations. Conceptual representations are, as we have seen above, representations whose functional role is to classify aspects of a scene, or items of a behavioral pattern, according to a certain category. This is the reason why only conceptual knowledge (whether verbally expressible or not) can enable overcoming the limits of situationally-bound use. Intuitively sampling objects, for example, on the basis of some salient similarity criterion, is a manifestation of knowledge-how, because it depends on situational features—for instance

that the situation represents some sort of average type to which the corresponding behavior is adapted. To overcome such situational limitations, categorical distinctions have to be introduced that enable the subject to transfer his or her knowledge partly to new situations that deviate, for instance, with respect to the objects that have been treated in standard situations. For example, a waiter who starts to work in a new restaurant using only coffee cups of one type, that is slightly higher than the large type used in the former restaurant, might fail in balancing the cup as long as he only takes recourse to his knowledge-how; but he might be more successful if he relied on a conceptual understanding of a distinguished large-cup-technique. In the same way, anticipation of the flight of a javelin is a situation-bound ability, since it depends on relatively rigid processing of visual information and proprioceptive mechanisms that are well-adapted to a range of standard cases, but fail for cases outside that range. If the case is exceptional (e.g., strong wind from behind), the subject can only attain success by analyzing the influence that this particular external condition will have on the standard performance. The same applies for knowledge-how expert chess knowledge, which fails in cases of random constellations because the experts' expertise in evaluating the scene is dependent on average situational features. The occurrence of "new" constellations requires extracting general properties from the scene, and thus has to be done by means of conceptual representations.

2. *Impenetrability versus penetrability of knowledge* is a contrast almost built into the notion of conceptuality that we propose. Non-conceptual representations are non-receptive for additional stimuli that could yield classificatory behavior. They have to be non-receptive ("impenetrable") in order to avoid interferences that could disrupt the more or less rigid mechanisms by which some well-defined type of behavior is regularly produced. Impenetrable knowledge-how, for example, is manifested by navigating ants cal-

culating their way home according to some rigid computational processes that are deployed on the basis of a small number of parameters. If the experimenter interferes with the process by repositioning the ant, the mechanism still works as it would have done without relocation, with the result that the ant misses the nest by exactly the distance and direction to which it has been repositioned by the experimenter (see Bartels & May 2009). In contrast, conceptually-based processing has to be penetrable in order to guarantee that categorical information can be extracted from the scene according to specific stimuli (in this case the repositioning stimuli) and used in evaluating the result produced by rigid processing up to the time of repositioning.

3. *Implicit versus explicit knowledge*. This distinction refers to whether or not the knowing organism has knowledge of the *rules* governing its knowledge application. For example, people learn the grammar of their natural language or internalize their society's norms implicitly, that is without knowledge of the principles that guide their language use or their social behavior. In such cases people represent rules only *indirectly*, by means of dispositions to have their reactions determined by the linguistic or social information in a way that can be recognized by their fellow subjects as to be in accordance with the rules. In contrast, explicit knowledge requires *direct* representation of rules, objects, or properties. The waiter in the restaurant, for example, after having achieved knowledge—that about his balancing of coffee cups, is able to refer directly to two sorts of cup shape, the high and cylindrical or the flat and bowl-shaped, respectively. In other words, he must be able to represent properties; if so, the waiter would, for instance, be able to draw inferences from the contents of his knowledge. Now, a person's ability to produce attribute-representations of objects presupposes the ability to apply *categories* to his or her own experience. For example, the waiter is able to represent coffee cups as high

and cylindrical objects because his capacities include the ability to apply the category of shape to the objects he is balancing. Thus, a person's possession of *conceptual capacities* is a condition that has to be fulfilled for his or her knowledge to be explicit. Moreover, given that the additional conditions for *conscious* processing of cognitive representations are fulfilled, the subject would then be able to consciously think about and to draw conscious inferences about the objects. In addition, *verbalizability* of knowledge depends on the presence of this conscious form of explicit knowledge.

4. *Automatic versus effortful processing.* As we have argued in (B), conceptuality entails openness to penetration. Now, if cognitive processing is receptive to penetration, additional costs in terms of attention and additional processing necessarily occur. If the ant's navigation mechanisms were receptive to a certain type of repositioning, it would have to use additional computational pathways for processing "repositioning information" and would be in need of additional calculation to determine the influence of the particular repositioning on the result produced by rigid calculation of the expected path back home.
5. *Continuous versus discontinuous processing.* Knowledge-that is characteristically used in a step-by-step manner with intermediate knowledge states (discontinuous), whereas knowledge-how appears to be grounded in smooth and fluent processing without intermediate states (continuous). The difference can be accounted for by the fact that knowledge-that is grounded in concept-based processing allowing for and instantiating discrete inferential steps, whereas knowledge-how is based on concept-free processing without clearly-defined intermediate knowledge states. An observable consequence of the continuous nature of knowledge-how is that lapses in knowledge use result in graded errors, or continuous distributions of errors (e.g., gradual precision losses of sensorimotor movements),

while lapses in use of knowledge-that express themselves in categorical errors, or discontinuous error distributions (e.g., switches of categories or total failures to come up with a result).

It is beyond the scope of the present article to give an outline of a research agenda for empirically confirming and underpinning the present account of knowledge-how compared to knowledge-that. Different examples of potential research areas and experimental paradigms have been pointed out in the preceding sections (e.g., numerosity judgments, spatial memory, intuitive knowledge use). The most convincing way to support the adequacy of the conceptuality criterion for distinguishing between knowledge-how and knowledge-that will be to run new experiments in these or other research areas that reveal behavioral and/or neural dissociations that comply with the distinction between concept-driven vs. concept-free knowledge-use along the lines of the different peculiarities of practical knowledge outlined above.

8 Conclusion

We have shown that *propositionality* is, in none of its three main senses, an adequate and useful demarcation criterion between knowledge-how and knowledge-that.

First, in its *semantic* sense (e.g., Stanley 2011a), propositionality applies to both knowledge-how and knowledge-that, and thus *a fortiori* cannot be successfully used as a demarcation criterion.

Second, in its "*language of mind*"-sense, propositionality applies to knowledge representation. As we have shown, the way in which a particular piece of knowledge is represented is independent from the type of knowledge exemplified by this piece of knowledge. Thus, again, this sense of propositionality is not useful as a demarcation criterion.

Third, propositionality in the sense of *linguistic, consciously available propositions* is without doubt a central phenomenological trait of knowledge-that as opposed to knowledge-

how. On the one hand, this sort of propositionality offers a rather trivial demarcation criterion. On the other hand, as a mere replication of a well-known phenomenological distinction, it can in no way be used to *explain* the different peculiarities characteristic of knowledge-how versus knowledge-that. *Anti-intellectualists* have tried to fill the void corresponding to non-propositionality, according to this third sense of propositionality, by declaring specific knowledge formats such as *sensorimotor or image-like* knowledge (Newen & Jung 2011). In our view, it is doubtful whether, with such an eclectic way of characterizing knowledge-how, a satisfactory and complete classification of knowledge-how could be achieved. We have, for example, argued that “intuitive” knowledge would be a further legitimate candidate for the list, and that it is, in all probability, not the only further candidate. Identifying different forms of knowledge-how without any well-grounded theoretical basis for the different forms will probably be of limited use for empirical research in cognitive science, neuroscience, and psychology.

In sum, “propositionality” can in none of its different senses provide a useful demarcation criterion for an empirically-fruitful theory of knowledge-how. Therefore, we go with the *intellectualists*, at least with respect to rejecting the propositionality criterion, but we depart where intellectualists fail to provide positive accounts of the obvious phenomenological and empirical peculiarities making knowledge-how distinct from knowledge-that. In contrast to the intellectualist position, we have provided a minimal notion of *conceptuality* as an alternative demarcation criterion. We suggest that conceptuality gives a sound basis for a fruitful theory of knowledge-how, and we have tried to provide support to this suggestion by showing that by means of an adequate notion of conceptuality, five central peculiarities of knowledge-how as compared to knowledge-that can be accounted for. Future research will have to show whether the framework for practical knowledge described here fulfills the empirical promise we think it has.

References

- Adams, M. P. (2009). Empirical evidence and the knowledge-that/knowledge-how distinction. *Synthese*, 170 (1), 97-114. [10.1007/s11229-008-9349-z](https://doi.org/10.1007/s11229-008-9349-z)
- Allen, C. (1999). Animal concepts revisited: The use of self-monitoring as an empirical approach. *Erkenntnis*, 51 (1), 33-40. [10.1023/A:1005545425672](https://doi.org/10.1023/A:1005545425672)
- Allen, C. & Hauser, M. D. (1991). Concept attribution in non-human animals: Theoretical and methodological problems in ascribing complex mental processes. *Philosophy of Science*, 58 (2), 221-240. [10.1086/289613](https://doi.org/10.1086/289613)
- Barch, D. M., Carter, C. S., Braver, T. S., Sabb, F., MacDonald, A., Noll, D. & Cohen, J. (2001). Selective deficits in prefrontal cortex function in medication-naïve patients with schizophrenia. *Archives of General Psychiatry*, 58 (3), 280-288. [10.1001/archpsyc.58.3.280](https://doi.org/10.1001/archpsyc.58.3.280)
- Bargh, J. A. & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist*, 54 (7), 462-479. [10.1037/0003-066X.54.7.462](https://doi.org/10.1037/0003-066X.54.7.462)
- Bartels, A. & May, M. (2009). Functional role theories of representation and content explanation: With a case study from spatial cognition. *Cognitive Processing*, 10 (1), 63-75. [10.1007/s10339-008-0226-y](https://doi.org/10.1007/s10339-008-0226-y)
- Bechara, A., Damasio, H., Tranel, D. & Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science*, 275 (5304), 1293-1295. [10.1126/science.275.5304.1293](https://doi.org/10.1126/science.275.5304.1293)
- Bechara, A., Tranel, D. & Damasio, H. (2000). Characterization of the decision making deficit of patients with ventromedial prefrontal cortex lesions. *Brain*, 123 (11), 2189-2202. [10.1093/brain/123.11.2189](https://doi.org/10.1093/brain/123.11.2189)
- Clancey, W. J. (1997). *Situated cognition: On human knowledge and computer representations*. Cambridge, MA: Cambridge University Press.
- Clark, A. (1997). *Being there: Putting brain, body and world together again*. Cambridge, MA: MIT Press.
- Devitt, M. (2011). Methodology and the nature of knowing how. *Journal of Philosophy*, 108 (4), 205-218.
- Dijksterhuis, A. & Nordgren, L. F. (2006). A theory of unconscious thought. *Perspectives on Psychological Science*, 1 (2), 95-109. [10.1111/j.1745-6916.2006.00007.x](https://doi.org/10.1111/j.1745-6916.2006.00007.x)
- Dreyfus, H. (2007). The return of the myth of the mental. *Inquiry*, 50 (4), 352-365. [10.1080/00201740701489245](https://doi.org/10.1080/00201740701489245)
- Dreyfus, H. & Dreyfus, S. (1986). *Mind over machine. The power of human intuition and expertise in the age of the computer*. Oxford, UK: Blackwell.
- Evans, G. (1982). *The varieties of reference*. Oxford, UK: Oxford University Press.

- Fodor, J. (1968). The appeal to tacit knowledge in psychological explanation. *Journal of Philosophy*, 65 (20), 627-640.
- Halligan, P. W., Kischka, U. & Marshall, J. C. (Eds.) (2004). *Handbook of clinical neuropsychology*. Oxford, UK: Oxford University Press.
- Hasher, L. & Zacks, R. T. (1979). Automatic and effortful processing in memory. *Journal of Experimental Psychology: General*, 108 (3), 356-388.
- May, M. & Klatzky, R. L. (2000). Path integration while ignoring irrelevant movement. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26 (1), 150-166. [10.1037/0278-7393.26.1.169](https://doi.org/10.1037/0278-7393.26.1.169)
- Milner, A. D. & Goodale, M. A. (1995). *The visual brain in action*. Oxford, UK: Oxford University Press.
- (2008). Two visual systems re-visited. *Neuropsychologia*, 46 (3), 774-785. [10.1016/j.neuropsychologia.2007.10.005](https://doi.org/10.1016/j.neuropsychologia.2007.10.005)
- Montello, D. R. (2005). Navigation. In P. Shah & A. Miyake (Eds.) *The Cambridge handbook of visuospatial thinking* (pp. 257-294). Cambridge, UK: Cambridge University Press.
- Myers, D. G. (2004). *Intuition. Its powers and perils*. New Haven, NJ: Yale University Press.
- Newcombe, N. S. & Huttenlocher, J. (2000). *Making space. The development of spatial representation and reasoning*. Cambridge, MA: MIT Press.
- Newen, A. & Bartels, A. (2007). Animal minds and the possession of concepts. *Philosophical Psychology*, 20 (3), 283-308. [10.1080/09515080701358096](https://doi.org/10.1080/09515080701358096)
- Newen, A. & Jung, E. M. (2011). Understanding knowledge in a new framework: Against intellectualism as a semantic analysis and an analysis of mind. In A. Newen, A. Bartels & E. M. Jung (Eds.) *Knowledge and representation* (pp. 79-105). Paderborn, GER: Mentis.
- Noë, A. (2005). Against intellectualism. *Analysis*, 65 (4), 278-290. [10.1093/analys/65.4.278](https://doi.org/10.1093/analys/65.4.278)
- Pepperberg, I. (1999). *The Alex studies*. Cambridge, MA: Harvard University Press.
- Pylyshyn, Z. W. (1984). *Computation and cognition*. Cambridge, MA: MIT Press.
- (1990). Computation and cognition: Issues in the foundations of cognitive science. In J. L. Garfield (Ed.) *Foundations of cognitive science: The essential readings* (pp. 18-74). New York, NY: Paragon House.
- Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*, 118 (3), 219-235. [10.1037/0096-3445.118.3.219](https://doi.org/10.1037/0096-3445.118.3.219)
- Ryle, G. (1949). *The concept of mind*. Chicago, IL: The University of Chicago Press.
- Snowdon, P. (2004). Knowing how and knowing that: A distinction reconsidered. *Proceedings of the Aristotelian Society*, 104 (1), 1-29. [10.1111/1467-9264.t01-1-00001](https://doi.org/10.1111/1467-9264.t01-1-00001)
- Spelke, E. S. (2000). Core knowledge. *American Psychologist*, 55 (11), 1233-1243. [10.1037/0003-066X.55.11.1233](https://doi.org/10.1037/0003-066X.55.11.1233)
- Spivey, M. (2008). *Continuity of mind*. Oxford, UK: Oxford University Press.
- Stanley, J. (2011a). Intellectualism and the language of thought: A reply to Roth and Cummins. In A. Newen, A. Bartels & E. M. Jung (Eds.) *Knowledge and representation* (pp. 41-49). Paderborn, GER: Mentis.
- (2011b). *Know how*. Oxford, UK: Oxford University Press.
- Stanley, J. & Williamson, T. (2001). Knowing how. *Journal of Philosophy*, 98 (8), 411-444.
- Stuss, D. T. & Alexander, M. P. (2007). Is there a dysexecutive syndrome? *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 362 (1481), 901-915. [10.1098/rstb.2007.2096](https://doi.org/10.1098/rstb.2007.2096)
- Toribio, J. (2008). How do we know how? *Philosophical Explorations*, 11 (1), 39-52. [10.1080/13869790701599044](https://doi.org/10.1080/13869790701599044)
- Tulving, E. & Schacter, D. I. (1990). Priming and human memory systems. *Science*, 247 (4940), 301-306. [10.1126/science.2296719](https://doi.org/10.1126/science.2296719)
- Young, G. (2011). Irreducible forms of knowledge-how in patients with visuomotor pathologies: An argument against intellectualism. In A. Newen, A. Bartels & E. M. Jung (Eds.) *Knowledge and representation* (pp. 51-77). Paderborn, GER: Mentis.

The Semantic Reading of Propositionality and Its Relation to Cognitive-Representational Explanations

A Commentary on Andreas Bartels & Mark May

Ramiro Glauer

Bartels and May propose an explanation of the difference between practical and theoretical knowledge in terms of the involvement of non-conceptual and conceptual representations, respectively. They thereby want to alleviate a shortcoming of Stanley's intellectualist theory of knowledge-how that cannot explain this difference. In this paper it is argued that an appreciation of the fact that both Stanley and Bartels and May employ a semantic reading of propositionality makes clear that their endeavors follow quite different goals. While Stanley gives an analysis of how we talk about knowledge-how, Bartels and May are interested in underlying cognitive representations. From Stanley's analysis of knowledge-how, nothing can be inferred about cognitive representations. The semantic reading of propositionality is then spelled out with the help of the idea that ascriptions of propositional attitudes are (like) measurement statements. Some considerations from measurement theory show how propositions can be used to reason about psychological states without themselves having to play any role in a person's psychology.

Keywords

Anti-intellectualism | Concepts | Conceptual representations | Homomorphic mapping | Intellectualism | Knowledge-how | Measurement | Measurement theory | Measurement view | Mental representation | Non-conceptual representations | Personal level | Propositional attitudes | Propositionality | Propositions | Semantic reading of propositionality

Commentator

[Ramiro Glauer](#)

ramiro.glauer@ovgu.de

Otto-von-Guericke-Universität
Magdeburg, Germany

Target Authors

[Andreas Bartels](#)

andreas.bartels@uni-bonn.de

Rheinische Friedrich-Wilhelms-Universität
Bonn, Germany

[Mark May](#)

mm@hsu-hh.de

Helmut-Schmidt-Universität
Hamburg, Germany

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

Bartels and May's paper presents the outlines of a theory of practical knowledge. The paper consists of a discussion of intellectualist and anti-intellectualist approaches to knowledge-how, a characterization of a range of behavioral particularities of practical knowledge, and the outlines of a theory that attempts to explain these behavioral particularities in terms of involved underlying mental representations. The discussion is remarkably clear, and the explicit exposition of what is to be explained by a theory of practical knowledge is a great virtue of the paper. For our purposes here, a discussion of the initial characterization of practical knowledge and its attempted explanation in terms of conceptual and non-conceptual capacities would help us assess the import of this paper. To my valuation, however, the discussion also reveals some very important features of the relation between knowledge ascriptions (and, to that effect, ascriptions of propositional attitudes in general) and descriptions of underlying cognitive structures and representations. Most importantly, Bartels and May employ Stanley's semantic reading of propositionality, according to which the propositionality of some mental state depends on whether a proposition is mentioned in the ascription of that state. As a result, questions concerning cognitive structure and underlying representations are largely detached from considerations concerning ascriptions of propositional attitudes. I think this is a great advantage, because we are not led to read back the relational grammatical structure of ascriptions of propositional attitudes onto psychological states themselves. Here I want to focus on this semantic reading of propositionality and ask about its effects on the relation between Bartels and May's proposed explanation of practical knowledge and Stanley's theory of knowledge-how. The result will be that Stanley and Bartels and May attempt to explain quite different things. While Stanley proposes a theory of how we ascribe knowledge-how to each other, Bartels and May are interested in underlying cognitive processes. The semantic reading of propositionality, however, only goes halfway towards disen-

tangling these different endeavors. A further step can be made with the help of the idea that ascriptions of propositional attitudes are (like) measurements. I will call this the *measurement view* of ascriptions of propositional attitudes. Considerations from measurement theory can then be used to shed further light on the relation between ascriptions of propositional attitudes and the underlying cognitive representations. The result will be that nothing can be inferred about cognitive structure from the structure of ascriptions of propositional attitudes alone. Propositions need not play any role in a theory of cognition. Nonetheless, there is a clear sense in which propositional attitudes are real. They are the measurement-theoretic representatives of behaviorally relevant states. In closing I will note that, given the close connection between concepts and propositions, a semantic reading of conceptuality might be desirable. For Bartels and May, this would mean that the difference between practical and theoretical knowledge should not depend on the conceptuality of the underlying representations. But given their definition of conceptuality, this would merely require a change in nomenclature.

Before going into the discussion of a semantic reading of propositionality, of measurement and its bearing on the relation between Bartels and May's proposed explanation of practical knowledge and Stanley's theory of knowledge-how, I will briefly summarize Bartels and May's line of argument.

2 The semantic reading of propositionality and the explanation of practical knowledge

Bartels and May set out to clarify what a theory of knowledge-how should provide and begin to give the outlines of such a theory. In their view, a theory of knowledge-how should explain the difference between practical and theoretical knowledge, the former being characterized by a number of distinguishing features. The proposal, then, is to explain this difference in terms of the reliance on non-conceptual capacities (or repres-

entations) in the case of practical knowledge and on conceptual capacities in the case of theoretical knowledge, instead of using propositionality as the main criterion. Their account of what is to be captured by a theory of know-how, and their proposed solution, are preceded by an illuminating discussion of the shortcomings of each side of the intellectualism vs. anti-intellectualism debate.

2.1 Merits and shortcomings of intellectualism

In short, Bartels and May claim that the intellectualists are right to concede that the distinction between knowing-how and knowing-that cannot be made in terms of the propositionality of knowing-that. Three readings of propositionality are distinguished:

- a representational reading, according to which the propositionality of some mental state depends on a sentence-like mental representation being tokened,
- a conscious-availability reading, according to which propositional representations are consciously available and can be expressed linguistically, and
- a semantic reading of propositionality, according to which the propositionality of some mental state depends on whether it is attributed as a propositional attitude.

It is argued that all three readings of propositionality are inapt for making the distinction between practical and theoretical knowledge. I take it that both the representational reading and the conscious-availability reading are implausible for independent reasons—the representational reading presupposes a language of thought, while the conscious-accessibility reading can arguably be undermined by considering cases in which someone would be said to know something she need not be able to express verbally, in terms of the proposition in question (this might involve some non-obvious logical consequences of one’s occurrent beliefs). In addition, the semantic reading is what our best in-

tellectualist account of knowledge-how, namely Stanley’s, employs, and Bartels and May follow Stanley’s analysis here.

According to the semantic reading of propositionality, whether some psychological attitude is propositional depends on the semantics of the locutions used to ascribe such attitudes. And our best current theories of the semantics of knows-wh locutions—i.e., of locutions that involve the verb “know” and some question word such as “who”, “where”, “what”, “when”, or, to that effect, “how”—tells us that knowledge-how is propositional—just as knowledge-that is. But as a result, it is argued, intellectualists are not able to explain the respective peculiarities of practical and theoretical knowledge—both are propositional. This is identified as the major shortcoming of intellectualism.

2.2 Merits and shortcomings of anti-intellectualism

The anti-intellectualists, on the other hand, lack a systematic criterion for the distinction between knowledge-how and knowledge-that. The introduction of different kinds of knowledge, based on different representational formats, by some anti-intellectualists is taken to be *ad hoc* (e.g., image-based knowledge and sensorimotor knowledge by Jung & Newen 2011). It is not based on an independently identified set of underlying representational formats that would explain the characteristic behavioral differences. Instead, it merely attempts to find alleged mental representational formats that intuitively fit the distinction (cf. Bartels & May this collection, p. 7). Further arguments to the effect that intellectualism is a non-starter are ineffective against Stanley’s (2011) version of intellectualism (cf. Bartels & May this collection, pp. 10-11). An attack from Toribio (2008, reference taken from Bartels & May this collection) to the effect that Milner & Goodale’s patient DF (cf. Milner & Goodale 1995) could not possibly have propositional knowledge of how to put a card into a slot presupposes that knowledge-how involves a *conceptual* grasp of how something is done or of what is acted upon. Roughly, Toribio argues that DF does not have proposi-

tional knowledge of how to put the card into the slot because she cannot report on the orientation of the slot. But Stanley acknowledges that some propositional attitudes involve the *non-conceptual* grasp of relevant states of affairs. In the case of DF, this involves the non-conceptual grasp of the orientation of the slot (cf. Stanley 2011, p. 172).

As a result, neither intellectualists nor anti-intellectualists provide a satisfactory account of knowledge-how. But both get some things right. The intellectualist is right in taking both knowledge-that and knowledge-how to be propositional. And the anti-intellectualist is right in requiring an explanation of the difference between these two kinds of knowledge, presumably in terms of underlying cognitive structures or kinds of mental representation.

2.3 Non-conceptual capacities as an explanation of practical knowledge

Bartels and May, then, pick up on the idea that practical knowledge might involve non-conceptual capacities, while theoretical knowledge is conceptual. They list a number of received peculiarities of practical knowledge that are to be captured by a theory of practical knowledge. And it is proposed that these peculiarities are the same peculiarities that result from a reliance on non-conceptual representations. Among the differential features of practical knowledge are its being context-bound, implicit, and automatic and effortless. Non-conceptual capacities, it is argued, just have these features. The result is a position that is intellectualist in form, because all kinds of knowledge are propositional, but anti-intellectualist in spirit, as the distinction of practical vs. theoretical knowledge is maintained. Practical knowledge is not reduced to theoretical knowledge; rather, the former is a non-conceptual form of knowledge while the latter is conceptual.

One effect of drawing the distinction between practical and theoretical knowledge in terms of *conceptuality* is that Bartels and May must follow Stanley in accepting non-conceptual forms of propositional knowledge. Patient DF cannot report on the orientation of the slot, but

nevertheless she non-conceptually grasps its orientation such that she is able to put the card into the slot. Due to her successful performance, she is said to know how to put the card into the slot, making this particular form of knowledge-how non-conceptual. This somewhat departs from tradition, where concepts are usually taken to be the constituents of thoughts, while thoughts are likely understood in a Fregean way as the intensions of sentences, i.e., propositions. It makes sense, though, because propositionality is understood semantically while conceptuality is not. Whether some cognitive capacity is conceptual or non-conceptual is thought to depend upon the kind of mental representation involved.

3 Knowledge ascriptions and mental representations

3.1 Analyzing knowledge ascriptions vs. explaining cognitive capacities

Now, it's easy to believe that the whole debate around propositions, concepts, non-conceptual representations, and cognitive structure is highly convoluted and that it is difficult to properly disentangle the different issues that lie behind a larger number of related debates. One important distinction, I take it, which is not always properly made, is whether one is concerned with what *someone* does (the whole person) as opposed to what his or her *cognitive system* does. What happens between Stanley's and Bartels and May's discussion of kinds of knowledge, then, is a shift from a personal-level perspective to a level at which the cognitive system is described.

Stanley formulates a theory of knowledge-how on the basis of an analysis of ascriptions of knowledge-how. And the subject of clear cases of appropriate knowledge-how ascriptions are persons. Their brains (or whatever else might realize their cognitive systems) can at best derivatively be said to know how to do something. This is made especially clear in Stanley's analysis, according to which knowledge-how involves first-person thought (cf. 2011, Ch. 3). If someone knows how to do something he knows

that a certain way of doing something is a way in which he could do it himself. It is hard to see how someone's cognitive system could have this kind of first-person thought in a non-derivative way.

Bartels and May, on the other hand, want to explain the particularities of practical and theoretical knowledge in terms of the involved underlying representations. As they put it at the outset of their discussion, “‘Explaining’ here is rather to be understood as showing how the realization of necessary conditions for the possession of concepts coincide with those conditions that have to be fulfilled in order to achieve the step from practical to theoretical knowledge, each characterized by their respective peculiarities. In other words, we search for ‘how-possibly-explanations’ of the peculiarities of practical versus theoretical knowledge” (Bartels & May [this collection](#)). “How-possibly-explanation” is a term from mechanistic accounts of explanation that characterizes attempted mechanistic explanations that are not yet well corroborated by an independent identification of the components of the alleged mechanism. Bartels and May clearly appeal to structures underlying cognitive abilities. In addition, they employ a notion of concepts that is further developed in Newen and Bartels (Bartels & Newen 2007), where it is made clear that concepts are kinds of mental representations (cf. [ibid.](#), p. 284). Their interest thus lies in the differences between the cognitive architectural realization of practical and theoretical knowledge, not in the ascription conditions of kinds of knowledge to persons. And, as said, among the virtues of Bartels and May's paper is the clarity of the exposition of what is to be explained by a theory of practical knowledge in the first place: the behavioral or functional peculiarities of practical knowledge.

I understand that making a distinction between different endeavors in philosophy of mind in terms of personal vs. sub-personal level explanations is not always a particularly attractive way to go about the problem. The personal level brings with it a number of loaded presumptions, for instance, concerning the import of norms for action and belief. And I do

not want to claim that such a rich conception of persons is involved in Stanley's discussion. Nonetheless it should be clear that Stanley is not interested in what the brain does, what its functional architecture is, or on which states it operates. He is interested in knowledge-how. And knowledge-how is something *someone* has: it's personal-level at least in the parsimonious way that it is something we attribute to each other.

In realizing that Bartels and May are really interested in the structure of cognitive systems possessing practical knowledge it becomes clear why they come to a conclusion that seems to be diametrical to what some other participants in the knowledge-how debate suggest. Bengson & Moffett (2007), for instance, argue that knowing how to do something is a matter of having a guiding conception of the way in which the subject of knowledge-how is to perform an activity. This captures that action guided by knowledge-how is a form of intelligent action—as opposed to something done by reflex, mere habit, or rote. It is an intellectual achievement to know how to do something. Bengson & Moffett (2007) argue that knowing how to do something requires an understanding of the activity at hand, and that understanding, in turn, is equivalent to the reasonable mastery of the concept that guides the action. Understanding is clearly something *someone* has; it is not a trait of his or her cognitive system that might rather be said to enable or mediate such understanding.

While the discussion in Bengson & Moffett (2007) sticks to the vocabulary of intellectual appraisal employed in the Rylean treatment of the topic, Bartels and May take a cognitive-psychological approach to the matter. For them, concepts are kinds of mental representations that serve to explain why someone has some ability. The notion of understanding does not figure prominently in their account. The difference to Bengson and Moffett's account can thus be traced back to different notions of what a concept is, which result from an interest in different perspectives on knowledge-how. Bengson and Moffett are interested in the conditions under which someone can be said to know how to do something, while Bartels & May want to ex-

plain the cognitive-psychological difference between practical and theoretical knowledge. When we adopt a semantic reading of propositionality and follow Stanley's analysis of knowledge-how, it becomes clear that these are very different endeavors. A theory of knowledge-how involves an analysis of what it is to ascribe such knowledge to someone; it is an investigation of the semantics of knowledge-how ascriptions and of our ways of talking. An explanation of the difference between practical and theoretical knowledge, on the other hand, tells us how corresponding abilities are realized by the cognitive system in terms of the employed representations.

One of the great virtues of a semantic reading of propositionality, then, is that it liberates us from drawing conclusions concerning cognitive architecture from the structure of ascriptions of mental states to subjects. Given that whether some mental state is propositional depends on the form of its ascription, there is no need to assume that the cognitive states described as propositional have to fulfill very specific conditions as to their structure and content. The correctness conditions for ascriptions of knowledge-how need not make reference to cognitive-architectural features of the subject of the ascription. And according to Stanley's analysis they don't. A knowledge state that is ascribed as propositional to some subject need not have propositional content itself nor be in any way structured such as to provide a vehicle for a propositional content. Indeed, Stanley (cf. 2011, p. 159) claims to have shown that having propositional knowledge states is entirely compatible with even an anti-representational conception of the mental. Nonetheless, knowledge-how is taken to be behaviorally real and efficacious, since it is implicated in certain actions and allows for explanations and predictions of behavior. We will shortly see how this can be so.

The liberation from cognitive-architectural commitments is somewhat occluded by Stanley, however, when he writes that he is interested in the *nature* of knowledge-how and that "[d]iscussions of semantics are often in fact discussions of metaphysics, carried out in the formal mode"

(Stanley 2011, p. 144). This appears to imply that ascriptions of propositional attitudes are understood realistically, and this in turn seems to be possible only if we take such ascriptions to describe real relations among subjects and mental representations to have the propositional content in question. This is the main motivation for a representational theory of mind (cf. Fodor 1987). Thus, an investigation into the *nature* of knowledge-how that comes to the conclusion that knowledge-how is propositional seems to employ a representational reading of propositionality.

Fortunately, this strong form of correspondence between ascriptions of propositional attitudes and the mental states that are thus described is not the only way to take such ascriptions to describe real mental states. We are not condemned to instrumentalism by adopting a semantic reading of propositionality when we recognize that ascriptions of propositional attitudes might share their logical structure with measurement statements.

3.2 Saving realism about propositional attitudes while employing a semantic reading of propositionality: A measurement view

At least since the late seventies a number of researchers have argued that having a propositional attitude is not a matter of standing in a certain cognitive relation to an abstract object, i.e., some particular proposition, but that ascriptions of propositional attitudes describe (intrinsic) psychological states with the help of a domain of abstract representatives, i.e., the domain of propositions. Propositions play the same role in ascriptions of propositional attitudes as numbers play in measurement statements (cf. e.g., Churchland 1979; Davidson 2001; Beckermann 1996; Matthews 2007). Let's call this the measurement view of propositional attitudes.

According to the measurement view, ascriptions of propositional attitudes have a non-relational logical form. The attitude verb and its propositional complement together form a complex predicate that refers to an intrinsic

psychological property of the subject of the ascription. Thereby the difficulty that propositional attitudes must be understood as a relation between a subject and a proposition is avoided: they could just as well be properties of the subject. A weaker form of the measurement view is exhausted by this claim (cf. e.g., Churchland 1979; Davidson 2001).

A stronger form of the measurement view in addition holds that ascriptions of propositional attitudes really are measurements in the sense that a formal measurement theory can be formulated for propositional attitudes (Matthews 2007). And indeed a further investigation of the analogy between ordinary measurement statements and ascriptions of propositional attitudes reveals how abstract objects can be used to refer to causally efficacious properties of objects without themselves playing any causal role. A measurement theory shows that one formal structure, the so-called *empirical structure*, can be homomorphically mapped onto another formal structure, the *representational structure*, the empirical structure being a formal theory about the domain of objects of interest (cf. e.g., Krantz 1972). The details of this mapping determine what can be inferred about the empirical structure from the representational structure. In length measurement, for instance, ratios between numbers correspond to ratios between lengths of objects.

Propositional attitudes figure in the explanation and prediction of behavior. Thus, in the case of propositional attitudes, the empirical formal structure has to be a formal theory of, presumably, the psychological states that are causally involved in the production of behavior. The representational formal structure has to be an adequate formalization of the structure of propositions. Leaving open what the two structures eventually turn out to be, it is the stronger claim that ascriptions of propositional attitudes really are measurements that I want to endorse here. In particular, I take it that propositions are the elements of a representational structure of a measurement theory for propositional attitudes. Let us have a brief look at measurement theory.

In ordinary measurements, numerical scales are used to represent systems of certain

measurable properties like length or mass, for example. Numbers are assigned to objects in accordance to a (procedural) rule. Somewhat simplified, in the case of length or mass measurement, a unit element is defined, and the number of unit elements that need to be concatenated in a certain way such as to be of equal length or mass, respectively, as the object that is measured, are counted. For mass the concatenation might be a simple lumping-together in the pan of a scale, while for length measurements unit elements are aligned rectilinearly. The number assigned to an object is equal to the count of unit-elements required. These numbers can then be used to represent relations among objects that are measured in the same way, i.e., on the same scale. An object that takes the number two on some length scale, for instance, is shorter than one that is assigned the number three, and it takes two objects of length two to get a concatenated object of equal length to an object that was assigned the number four on that scale. Thus, the system of objects is mapped with respect to their length onto the formal structure constituted by the natural numbers, including addition and the less-or-equal relation. The result is a homomorphic mapping from objects to numbers that respects certain additive relations among the lengths of objects. Correspondingly, the addition of numbers can be used to reason about the lengths of objects. Other properties of these objects and their relations might not be captured by the homomorphism. Which numerical operations can be used to reason about the objects' properties of interest depends on the scale that is used. In temperature measurement, for instance, most common scales do not respect ratios among temperatures, such that it does not make sense to say, for instance, that the air on a sunny day at 28° centigrade is twice as warm as the air on a day in fall at 14° centigrade.

Importantly, the objects' properties of interest are *holistically* captured by the numbers on a scale. It is in virtue of their position on the scale and the admissible operations that numbers represent certain (amounts of) properties of measured objects. There is nothing intrinsic to the number five that would make it a repres-

entative of a length of five centimeters or a weight of five kilograms. Individually, i.e., without their position on a scale, numbers don't tell us anything about the property they are used to represent—not even when the dimension (length, weight, ...) is added. Thus, which numbers represent which property (or amount of a property) and which operations on these numbers can be used to reason about the property of interest depends on the employed scale. Neither are all relations among objects respected by the homomorphic mapping; nor can all relations among the numerical representatives be read back onto the objects of interest. This much can be said on the basis of basic measurement theory as formulated by [Krantz et al. \(1971\)](#).

Most interestingly for our present purposes, measurement in the sense of homomorphic mapping does not require numerical representatives. Elements of other abstract structures might just as well serve as the targets of such homomorphic mappings. This idea is exploited by [Matthews \(2007\)](#) and [Dresner \(2010\)](#), for instance. In particular, Matthews argues that the structure of propositions, including their inferential and evidential relations among each other and to perceptions, might thus serve as a measurement structure for certain psychological states of subjects: those that are commonly called the propositional attitudes. These psychological states are homomorphically mapped onto propositions—the causal relations among the former being captured by the inferential, and other relations among the latter. The propositions can then be used to identify psychological states and, importantly, to reason about them. Thereby, propositional attitudes can appear in explanations and predictions of behavior without the propositions themselves having to play any causal role in the cognitive system.

I take it that propositional structures represent psychological properties holistically—just as numerical structures represent properties of objects holistically. The homomorphic mapping as a whole respects certain relations among psychological states, and it is in virtue of their position within the propositional structure that

particular propositions can be said to represent some psychological state. According to this view, there is nothing intrinsic to propositions that would relate them to particular psychological states. Thus, a measurement-theoretic notion of propositionality does not require the states that are referred to with the help of propositions to have propositional content themselves. Nonetheless, ascriptions of propositional attitudes can be understood realistically just as ordinary measurements are understood realistically. Once the mapping is fixed, it is an entirely objective question which proposition represents some given psychological state.

Neither numbers nor propositions are themselves taken to be causally relevant, but they are used to pick out a particular causally relevant property (or state) from a range of possible relevant properties (or states) as defined by the scale in use. Numbers on a meter scale are used to identify the length of objects. And it is the length of a pole, say, that is relevant for building a rack, not the number that is used to identify that length. The number is only relevant in relation to the numbers that are assigned to other parts of the rack. Similarly, propositions are used to identify psychological states that are behaviorally relevant. But it is the psychological states themselves that produce behavior, not the propositions that are used to identify them. Using propositions to identify psychological states leaves open how these states are realized within the cognitive system. All that is required is that the homomorphism holds. Indeed, drawing conclusions about the structure of the cognitive system from observations concerning properties of the propositional representatives of psychological states that are not warranted by the representational scheme (or “scale”) arguably amounts to an over-assignment of structure (cf. [Dresner 2004](#)). As noted above, not all properties of the system of representatives are shared by what they represent. The homomorphism holds with respect to some structural features of the represented objects as determined by the used scale.

Stanley appears to be at least sympathetic to such a measurement-theoretic conception of propositions—he mentions [Matthews \(2007\)](#) ap-

provingly. And there is reason to believe that such a measurement account of ascriptions of propositional attitudes is a plausible candidate for a semantic conception of propositionality. As mentioned above, it has the advantage of giving a non-instrumentalist, realist account of propositional attitudes without buying into any direct correspondence between propositions and mental representations that would lead to a language-of-thought-like theory of cognition. While Fodorean Realism presupposes that ascriptions of propositional attitudes can only be correct if the involved terms refer to actual cognitive entities and relations (i.e., a functional/computational relation towards a mental representation, where the former determines the kind of attitude and the latter its propositional content), such a measurement account makes clear how a system of propositions could structurally (i.e., holistically) represent psychological states without having to assume that psychological states themselves have propositional content or, at any rate, are dependent on how they are ascribed. And it eschews some of the difficulties associated with more traditional accounts, such as explaining how propositions can both be the abstract, sharable contents of thoughts and at the same time psychologically real in that what someone does depends on the contents of his desires and beliefs, etc. (cf. Davidson 2001). The mental states represented by some propositional attitude ascriptions are psychologically real; the proposition itself need not be. First of all, it serves as a representative for that state.

The difference between Stanley's and Bartels and May's accounts of knowledge-how and practical knowledge, respectively, can then be understood as follows. Stanley is interested in the structure of the domain of abstract entities that are used to represent psychological structure, while Bartels and May are interested in the structure of the empirical domain of psychological entities and relations that are described in terms of propositional attitudes. Both endeavors are related in that they involve a phenomenon that we might call "knowing how to do something", and both use intuitive examples and empirical evidence as test cases for their accounts. But their respective goal is really quite

different. In analogy to the measurement of length, one might say that Bartels and May are interested in giving a theory of how different bodies behave with respect to their length under some range of (physical) concatenation operations and comparison relations. For instance, welding two rods might have an influence on the resultant length of the composite rod such that it is not equally long as the two aligned but unwelded rods. Or, they might be interested in how length measurement transfers to smaller scales, such as molecular, atomic, or subatomic distances. Stanley, on the other hand, would be interested in the more formal properties of the numerical scales that are used for length measurement. He might ask how different scales relate. Just as the Fahrenheit scale can be transferred into the centigrade scale, knows-wh locutions might be transformed into know-that locutions.

Toribio's above-mentioned attack on intellectualism would then not be successful, because she has not realized that Stanley's theory really is about the structure of the representatives of certain psychological states, and not about the psychological states themselves. She offers some considerations concerning the structure of the psychological states that are meant to show that they could not possibly be propositional. But she does not give us a reason to think that the considered properties of certain cognitive processes face difficulties in terms of being represented by a propositional structure. Stanley then shows that there is no such difficulty. Toribio's discussion, on the other hand, is rather interesting for the development of an account of the cognitive structures that make it the case that someone knows how to do something.

Stanley's and Bartels and May's accounts are thus relatively independent of each other. Stanley's theory of knowledge-how can be seen as a partial investigation of the representational structure that we use to identify certain mental states. The approach of Bartels and May, on the other hand, is an attempt to give an explanation of certain cognitive capacities that are taken to be expressions of knowledge-how in terms of underlying mental representations. Given that propositional attitude ascriptions

measure psychological states, they aim to formulate a theory of the empirical structure. The measurement view first of all serves to disentangle these different endeavors and to shed some light on the relation between them, namely that the search for underlying representations and mental mechanisms is largely unconstrained by the structure of ascriptions of propositional attitudes by themselves and that conclusions about the empirical structure can only be drawn when the mapping is known as well.

This take is in line with both Stanley's theory and Bartels and May's explanation of practical knowledge. Stanley believes that cognitive psychology does not decide whether knowledge-how is propositional and refutes all objections to the contrary. The propositionality of knowledge-how is a matter of the semantics of their ascriptions. And Bartels and May give a characterization of the difference between practical and theoretical knowledge that is independent of Stanley's theory of knowledge-how. Practical knowledge has some behavioral/functional characteristics that are to be explained in terms of mental representations. The measurement view parts company with Stanley in his contention that he provides an investigation into the *nature* of knowledge-how. Rather, the measurement view is an investigation into a part of the representational structure of a measurement theory for a certain range of psychological states. We would not take an investigation of the centigrade scale to be an investigation of the nature of temperature.

4 Some final remarks

What the discussion around knowledge-how mainly shows, I think, is that the relation between propositional attitudes, cognitive structures or representations, and the behavioral evidence for their respective presence are still not well understood. It seems that we find it surprisingly difficult to disentangle our different ways of talking about ourselves and others in terms of what we believe, on the one hand, and in terms of the information that our brains (or some other division of the body-environment) process on the other. The main difficulty seems

to be that we take ascriptions of propositional attitudes to mirror psychologically real relations between subjects and propositions. As such, we feel the need to tell a story about how propositional attitudes are realized in the brain. The measurement view enables us to employ a less committal way of representing someone's psychological states that largely leaves open how the cognitive system manages to coordinate its behavior with the environment. The constraints that are put on cognitive architecture by successful ascriptions of propositional attitudes are really quite weak. To be sure, if the measurement view is to be proven correct, there must be a homomorphic mapping from an empirical structure into the propositional structure. But homomorphisms abound. Any number of homomorphisms can be found between any two structures. And as far as we can tell, the structure of propositions is homomorphic to the course of the sun and the stars. This is why we can employ intentional explanations for just about any system we want. The measurement view becomes informative when we have formalizations of the two structures and a measurement theory that describes the particular homomorphism of interest that holds between them. Then we can tell what we learn about the empirical structure by means of reasoning about propositions. An attempt to infer the empirical structure from the representational structure alone must fail.

In the case of propositional attitudes, I ultimately doubt that the mapping is best conceived as holding between internal cognitive architectural structure and propositional attitude ascriptions. Propositional attitudes might rather be measurements of structures of observable behavior. Propositional attitudes are ascribed on the basis of observable behavior together with some standards of folk psychology—such as that one believes what one sees or what one is told by trustworthy peers. Propositions might provide standardized ways of identifying behaviorally relevant circumstances, including what someone saw, was told, and aims for, that would otherwise have to be identified less systematically by way of particular situations and individual histories. I can tell that you know that the earth is an approximate sphere—you've

certainly learned it somewhere. I do not need to go back in your learning history until I find the moment in which someone uttered a sentence with the respective meaning—which would allow for similar predictions and explanations.

Taking propositional attitude ascriptions to be measurements of structures of observable behavior would also be very much in line with Ryle's original, rather behaviorist discussion of knowledge-how. With reference to our use of mental vocabulary to describe the behavior of others, Ryle writes that "we go beyond what we see them do and hear them say, but this going beyond is not a going behind, in the sense of making inferences to occult causes; it is going beyond in the sense of considering, in the first instance, the powers and propensities of which their actions are exercises" (1949, p.51). The powers and propensities are in turn understood as complex dispositions, describable in terms of their acquisition and manifestation conditions. The move from a structure of observable behavior to a propositional structure would take the place of acknowledging the role of so-called internal states; for now we can exploit inferential relations among propositions for explanation and prediction. But these propositional attitudes need not be understood as internal states. Instead they could be taken as measurement representations of Ryle's powers and propensities. Ryle notwithstanding, however, we need not give up cognitive psychology. Ascriptions of propositional attitudes and cognitive representations would relate via the behavior that each is to explain—they provide complementary explanations of the same behavior. For Bartels and May's explanation of practical knowledge this would mean that it is not part of a theory of an empirical structure for measurements of propositional attitudes. It would be a cognitive-psychological explanation of a behaviorally characterized psychological phenomenon called practical knowledge. The main point of this commentary, though—namely, that Stanley and Bartels and May are up to different things and that little can be inferred about cognitive architecture from Stanley's analysis of knowledge-how—remains untouched.

In closing, I want to mention one reservation that can be held against the particular cognitive-architectural account presented by Bartels and May. Given that concepts remain a vexed issue in contemporary discussion, that they are traditionally closely related to propositions, and that it is notoriously difficult to find good grounds for attributing representations of a certain kind and with a specific content to cognitive systems that are not able to verbally express their beliefs, a *semantic reading of conceptuality* might be worth considering. Concepts might be broadly conceived of as the constituents of thoughts, i.e., (trains) of propositional attitudes. In our case: whatever is a constituent of knowledge-how would count as a concept. One effect of this would be that the reliance on non-conceptual capacities in order to explain certain forms of knowledge-how, like that of patient DF, would not be open to Stanley. But as an alternative, Stanley could accept demonstrative concepts and claim that some forms of knowledge-how are distinguished by their involvement. Admittedly, Bartels & May would have to change their terminology; their abilities approach to concepts is not compatible with concepts being the constituents of propositions alongside a semantic reading of propositionality. But nothing much seems to be lost by this. Quite possibly, mentalistic vocabulary is just not the best way to come to grips with the structure of cognitive systems.

References

- Bartels, A. & May, M. (2015). What a theory of knowledge-how should explain. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Bartels, A. & Newen, A. (2007). Animal minds and the possession of concepts. *Philosophical Psychology*, 20 (3), 283-308. [10.1080/09515080701358096](https://doi.org/10.1080/09515080701358096)
- Beckermann, A. (1996). Is there a problem about intentionality? *Erkenntnis*, 45 (1), 1-23. [10.1007/BF00226368](https://doi.org/10.1007/BF00226368)
- Bengson, J. & Moffett, M. A. (2007). Know-how and concept possession. *Philosophical Studies*, 136 (1), 31-57. [10.1007/s11098-007-9146-4](https://doi.org/10.1007/s11098-007-9146-4)
- Churchland, P. M. (1979). *Scientific realism and the plasticity of mind*. Cambridge, UK: Cambridge University Press.
- Davidson, D. (2001). What is present to the mind. In D. Davidson (Ed.) *Subjective, intersubjective, objective* (pp. 53-68). Oxford, UK: Oxford University Press.
- Dresner, E. (2004). Over-assignment of structure. *Journal of Philosophical Logic*, 33 (5), 467-480. [10.1023/B:LOGI.0000046068.00813.83](https://doi.org/10.1023/B:LOGI.0000046068.00813.83)
- (2010). Language and the measure of mind. *Mind & Language*, 25 (4), 418-439. Blackwell Publishing Ltd. [10.1111/j.1468-0017.2010.01396.x](https://doi.org/10.1111/j.1468-0017.2010.01396.x)
- Fodor, J. (1987). The persistence of the attitudes. In J. Fodor (Ed.) *Psychosemantics* (pp. 1-26). Cambridge, MA: MIT Press.
- Jung, E.-M. & Newen, A. (2011). Understanding knowledge in a new framework: Against intellectualism as a semantic analysis and an analysis of mind. In A. Newen, A. Bartels & E.-M. Jung (Eds.) *Knowledge and representation* (pp. 79-105). Stanford, CA: Centre for the Study of Language & Information.
- Krantz, D. H. (1972). Measurement structures and psychological laws. *Science, New Series*, 175 (4029), 1427-1435. [10.1126/science.175.4029.1427](https://doi.org/10.1126/science.175.4029.1427)
- Krantz, D., Luce, D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement*. New York, NY: Academic Press.
- Matthews, R. J. (2007). *The measure of mind*. Oxford, UK: Oxford University Press.
- Milner, A. D. & Goodale, M. A. (1995). *The visual brain in action*. Oxford, UK: Oxford University Press.
- Ryle, G. (1949). *The concept of mind*. Chicago, IL: University of Chicago Press.
- Stanley, J. (2011). *Know how*. Oxford, UK: Oxford University Press.
- Toribio, J. (2008). How do we know how? *Philosophical Explorations*, 11 (1), 39-52. [10.1080/13869790701599044](https://doi.org/10.1080/13869790701599044)

Preparing the Ground for an Empirical Theory of Knowing-How

A Reply to Ramiro Glauer

Andreas Bartels & Mark May

The commentary gives a clear and instructive summary of our main arguments against both, intellectualist and anti-intellectualist accounts of knowing-how. But the aim of our account is not correctly described as an attempt to give an explanation of certain cognitive capacities that are taken to be expressions of knowing-how in terms of underlying mental representations. (Glauer [this collection](#), p.10). What we aim at is not an empirical theory of knowing-how, but a framework that would be useful for cognitive scientific research on phenomena of knowing-how.

Keywords

(Anti-) intellectualism | Conceptuality | Knowing-how | Knowing-that | Knowledge representation | Propositionality

Authors

[Andreas Bartels](#)

andreas.bartels@uni-bonn.de

Rheinische Friedrich-Wilhelms-Universität
Bonn, Germany

[Mark May](#)

mm@hsu-hh.de

Helmut-Schmidt-Universität
Hamburg, Germany

Commentator

[Ramiro Glauer](#)

ramiro.glauer@ovgu.de

Otto-von-Guericke-Universität
Magdeburg, Germany

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Answer to the Commentary

First, we want to thank Ramiro Glauer and emphasize that his commentary gives a clear and instructive summary of our main arguments against both intellectualist and anti-intellectualist accounts of knowing-how (see Section 2). As he rightly points out, we are parting ways with

[Jason Stanley \(2011\)](#) with respect to the issue of *propositionality* as an alleged demarcation criterion between knowing-how and knowing-that. There are at least three different conceptions of propositionality, and none turns out to be helpful in making the distinction. In particu-

lar, the semantic reading of propositionality, according to Stanley's thoughtful and impressive account, applies to clear-cut cases of knowing-how. Since knowing-how is no less propositional, according to the semantic reading, than knowing-that, there is no hope of understanding the peculiarities of knowing-how by adopting such a stance.

In Section 3, Glauer then turns to what in his opinion is the main difference between Stanley's and our account. Unfortunately, we don't think that he quite grasps the point that is important to us when he argues that "what happens between Stanley and Bartels & May's discussion of kinds of knowledge, then, is a shift from a personal-level perspective to a level at which the cognitive system is described" (Glauer [this collection](#), p. 4), and later, "Bartels & May, on the other hand, want to explain the peculiarities of practical and theoretical knowledge in terms of the involved underlying representations" (Glauer [this collection](#), p. 5). This, we have to say, is clearly a misrepresentation of our account and the intentions behind our developing it.

To be more specific, we argue that neither the semantic nor the *representational* reading of propositionality is suited to grounding the distinction between knowing-how and knowing-that (Bartels & May [this collection](#), pp. 5–6): "[w]hether a piece of knowledge is a case of practical or of theoretical knowledge does not depend on whether it is supported by language-like structures or not" (p. 6). Thus, contrary to the picture drawn in the commentary, we agree with Stanley with respect to his denial of a *representational* demarcation criterion between knowing-how and knowing-that. We thereby don't want to express any anti-representational reservations (as is also the case, in our opinion, for Stanley). However, we are skeptical with respect to any type of account that, in rather intuitive ways, identifies kinds of knowledge with ways of representing knowledge. This indeed is our main issue of disagreement with the anti-intellectualists (Glauer mentions this on p. 3).

What about the "shift from a personal-level perspective to a level at which the cognitive system is described" that Glauer mentions

([this collection](#), p. 4)? First, we are not quite sure how Glauer would himself mark the difference between a "person" and a "cognitive system", and what relevance he would ascribe to that difference with respect to the issue of knowing-how. Our paper wants to make clear that the *first-person-perspective* is an important constituent in the analysis of the specific dispositional states that characterize "practical ways of thinking"—specific ways of epistemic access to propositional contents when knowing-how is at stake (Bartels & May [this collection](#), p. 6). Thus, we agree that the knowing person, including all of his or her cognitive capacities and behavioral resources, has to be taken into account for a thorough analysis of knowing-how; see, for instance, our example of the waiter in a restaurant balancing different types of coffee cups (p. 16).

In essence, Ramiro Glauer's commentary draws a picture of our account that misses its main intentions. The aim of our account is not correctly described as "an attempt to give an explanation of certain cognitive capacities that are taken to be expressions of knowledge-how in terms of underlying mental representations" (Glauer [this collection](#), p. 9). Instead, our aim is to identify and specify some constituents of an empirically fruitful theory of knowing-how. In a first step, as we argue, this requires a careful description of central epistemic peculiarities that characterize knowing-how *as opposed to* knowing-that, and that thus have to be covered by any adequate theory (see Bartels & May [this collection](#), pp. 12–13). We then ask what general sort of *epistemic capacities* may coincide with the peculiar capacities embodied by knowing-how and knowing-that, respectively. And finally, we suggest that *conceptuality* versus *non-conceptuality* may be the general distinction that coincides with typical knowing-that and knowing-how-capacities, and go on to highlight some of the explanatory virtues of such a proposal. For the last step we use a theory that characterizes conceptual abilities by specific behavioral traits (Newen & Bartels 2007).

Our approach to the problem leaves open by what types of *mental representations* those conceptual abilities may be supported, if at all.

It cannot even be guaranteed that the distinctions drawn within our conceptual framework coincide with any distinctions between representational formats. What we aim at is not an empirical theory of knowing-how, but a *framework* that would be useful for cognitive scientific research on the phenomena of knowing-how. Thus, it may turn out to be useful to fill that framework with psychological or neurological hypotheses concerning representational mechanisms that may produce the epistemic capacities characterizing knowing-how. In Section 7 of our paper (Bartels & May this collection, pp. 16–17) we have provided different empirical examples of mainly psychological research that has already been undertaken in this line.

We are looking at the subject not so much from the perspective of philosophers of mind, but from the perspectives of philosophy of science and psychology. We therefore do not see good reasons to go into any detail of the specific theory that Ramiro Glauer explores in the second part of his commentary (this collection, pp. 6–7), namely the *measurement view* of propositional attitudes (Matthews 2007). Since our contribution does not intend to propose a new theory of knowing-how, it would be quite pointless to compare the potential merits of such a theoretical view with our own account. What we suggest is that psychological research, or cognitive scientific research more generally, may work along the path we have outlined, and thus make progress in explaining knowing-how.

2 Conclusion

We agree to the commentary concerning our main arguments against both, intellectualist and anti-intellectualist accounts of knowing-how. But we disagree with it concerning the picture that it draws of the aim of our account.

References

- Bartels, A. & May, M. (2015). What a theory of knowledge-how should explain. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Glauer, R. (2015). The semantic reading of propositionality and its relation to cognitive-representational explanations: A commentary on Andreas Bartels & Mark May. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Matthews, R. J. (2007). *The measure of mind*. Oxford, UK: Oxford University Press.
- Newen, A. & Bartels, A. (2007). Animal minds and the possession of concepts. *Philosophical Psychology*, 20 (3), 283–308. [10.1080/09515080701358096](https://doi.org/10.1080/09515080701358096)
- Stanley, J. (2011). *Know how*. Oxford, UK: Oxford University Press.

Introspective Insecurity

Tim Bayne

This paper examines the case for pessimism concerning the trustworthiness of introspection. I begin with a brief examination of two arguments for introspective optimism, before turning in more detail to Eric Schwitzgebel's case for the view that introspective access to one's own phenomenal states is highly insecure. I argue that there are a number of ways in which Schwitzgebel's argument falls short of its stated aims. The paper concludes with a speculative proposal about why some types of phenomenal states appear to be more introspectively elusive than others.

Keywords

Cognitive phenomenology | Emotion | Freestanding judgments | Imagery | Introspection | Introspection-reliant | Optimism | Pessimism | Scaffolded judgments | Schwitzgebel

Author

Tim Bayne

tim.bayne@manchester.ac.uk

The University of Manchester
Manchester, United Kingdom

Commentator

Maximilian H. Engel

M.H.Engel.1@student.rug.nl

Rijksuniversiteit Groningen
Groningen, Netherlands

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

There is a curious ambivalence in current attitudes towards our epistemic relationship to consciousness. Some theorists hold an optimistic view of the powers of introspection, regarding judgments about one's current experiences as epistemically secure—perhaps some of the most secure judgments that we make. Optimists rarely claim that we have exhaustive and infallible access to consciousness, but they do hold the epistemic credentials of introspection in high regard, at least when introspection is directed towards the phenomenal character of consciousness. Those inclined to optimism don't doubt that it is possible to mis-remember or mis-report one's experiences, but they tend to assume that one has some

kind of epistemic access to one's experiences simply by having them.¹

Running alongside this vein of optimism is a rather more pessimistic strand of thought, according to which the epistemic credentials of introspection are chronically insecure. Far from regarding introspection as a light that illuminates every corner of consciousness, pessimists suspect that significant swathes of experience are accessible to introspection only with great difficulty if at all.²

¹ Theorists inclined towards optimism include Chalmers (2003), Gertler (2012), Goldman (2004), Horgan et al. (2006), Horgan & Kriegel (2007), Siewert (2007), and Smithies (2012).

² The contrast between "optimists" and "pessimists" is far from sharp, for optimists often grant that epistemic access to consciousness can be (very) challenging, and pessimists often allow that there are experiential domains with respect to which introspection is trust-

Glossary

Introspection	An unmediated judgment that has as its intentional object a current psychological or phenomenal state of one's own.
Discrimination	The capacity to attentively single the state out from amongst the other experiences that one has at the time in question.
Categorize	To categorize a phenomenal state is to locate it within a taxonomy of some kind.
Directly and indirectly introspective judgments	A direct introspective judgment concerns the phenomenal character/content of one's current phenomenal state(s) and is grounded in a single act of introspective attention, whereas an indirect introspective judgment concerns the general nature of one's conscious experience and is not grounded in a single act of introspective attention.
Scaffolded judgments	An introspective judgment is scaffolded if and only if it is accompanied by a disposition to make a first-order judgment (e.g., a perceptual judgment) whose content broadly corresponds to the judgment of the introspective judgment. For example, the judgment that one has a visual experience as of a red tomato in front of one is scaffolded insofar as it is accompanied by a disposition to make the perceptual judgment that there is a red tomato in front of one.
Freestanding judgments	An introspective judgment is freestanding if and only if it is not accompanied by a disposition to make a first-order judgment (e.g., a perceptual judgment) whose contents broadly corresponds to the judgment of the introspective judgment.

According to Dan Haybron, “[...]even the gross qualitative character of our conscious experience can elude our introspective capacities” (Haybron 2007, p. 415). Sounding a similar note, Maja Spener has argued that “philosophers and psychologists routinely overestimate the epistemic credentials of introspection in their theorizing” (Spener unpublished; see also Spener 2011a, 2011b, and 2013). But perhaps the most thoroughgoing pessimist is Eric Schwitzgebel:

Most people are poor introspectors of their own ongoing conscious experience. We fail not just in assessing the causes of our mental states or the processes underwriting them; and not just in our judgments about nonphenomenal mental states like traits, motives and skills, and not only when we are distracted, or passionate or inattentive or self-deceived, or pathologically deluded or when we're re-

worthy. Nonetheless, these terms are useful insofar as they capture the overarching attitude that the two groups of theorists express with regard to introspection.

flecting about minor matters, or about the past, or only for a moment, or when fine discrimination is required. We are both ignorant and prone to error. There are major lacunae in our self-knowledge that are not easily filled in, and we make gross, enduring mistakes about even the most basic features of our currently ongoing conscious experience (or “phenomenology”), even in favourable circumstances of careful reflection, with distressing regularity. (2008, p. 247)

Although Schwitzgebel's pessimism is tempered by moments of optimism, the dominant theme in his work is that introspection cannot be trusted to reveal anything other than the most mundane features of consciousness. Descartes, Schwitzgebel argues, “had it quite backwards when he said the mind—including especially current conscious experience—was better known than the outside world” (2008, p. 267).

I feel the pull of both optimism and pessimism. In my optimistic moments I find it hard

to take seriously the suggestion that I might be guilty of “gross and enduring mistakes” about the basic features of my current phenomenology. But the arguments for pessimism are powerful and not easily dismissed, and I worry that Schwitzgebel is right when he suggests that the allure of optimism might be due to nothing more than the fact that “no-one ever scolds us for getting it wrong” (2008, p. 260).

A central aim of this paper is to provide an overview of Schwitzgebel’s case for introspective pessimism, and to chart a number of ways in which the optimist might respond to it. But although this paper can be read as a defence of a kind of optimism, my central concern is not so much to take sides in this debate as to advance it by noting various complexities that have perhaps been overlooked. But before turning to the debate itself let me make a few comments about its importance. An account of the trustworthiness of introspection is likely to have a bearing on two important issues. Most obviously, it has implications for the use of introspection as a source of evidence regarding philosophical and scientific debates about consciousness. Whether or not introspection is our sole form of access to consciousness, there is no doubt that it is currently treated as a *central* form of such access, and thus doubts about the reliability of introspection engender doubts about the viability of the study of consciousness. A second issue on which the trustworthiness of introspection has an important bearing concerns debates about the nature of introspection, and in particular the relationship between introspection and consciousness. Some accounts of introspection take a person to be necessarily acquainted with his or her conscious states, where acquaintance is an epistemic relationship of a particularly intimate kind (Gertler 2012; Horgan et al. 2006; Smithies 2012). It is fair to say that such approaches are optimistic by nature, and although advocates of such accounts have attempted to accommodate the possibility of introspective ignorance and error (see e.g., Horgan 2012), the success of such attempts is very much an open question. Other accounts of “introspection”—such as those that deny that there are any distinctively first-per-

sonal modes of access to consciousness—can easily accommodate introspective ignorance and error, but they struggle to account for the epistemic security that often seems to characterize introspection. In short, an account of introspection’s epistemic profile would function as a useful constraint on accounts of its nature.

2 Motivating optimism

By “introspection” I mean an unmediated judgment that has as its intentional object a current psychological state of one’s own. Introspection can take as its object a wide variety of psychological states, but here I am concerned only with the introspection of *phenomenal states*—states that there is “something it is like” for the subject in question to be in. In principle one could have any number of reasons for self-ascribing a phenomenal state—for example, it is possible to self-ascribe pain on the basis of neural or behavioural evidence—but introspection involves the self-ascription of phenomenal states on the basis of seemingly “direct” contact with them.³

There are many aspects of consciousness with respect to which we clearly have little to no introspective access. For example, introspection is clearly not a source of information about the neural basis of consciousness or its functional role. But surely, one might think, introspection can provide trustworthy answers to such questions as, “Am I now in a conscious state with such-and-such a phenomenal character?” Roughly speaking, to regard introspection as able to reveal the phenomenal character of one’s conscious states is to have an optimistic attitude towards it. But there is more than one sense in which introspection might be said to reveal the character of consciousness, and thus more than one way to be an introspective optimist.

One way in which introspection can reveal a phenomenal state is by allowing one to *discriminate* it from its phenomenal neighbours. I take discrimination to be bound up with the ca-

³ Introspection may involve direct access to consciousness at a personal level and yet also be inferential and indirect at sub-personal levels of description.

capacity to single the state out from amongst the other experiences—e.g., thoughts, perceptual experiences, and bodily sensations—that happen to populate one’s field of consciousness. Discriminative access to an experience allows one to direct one’s attention towards it and to thus make it the potential target of demonstrative thought—“I wish that *this* experience would stop”. A second mode of introspective access to consciousness involves the deployment of categories. To categorize a phenomenal state is to locate it within a taxonomy of some kind. Categorical access to the experience of an itch, for example, involves recognizing it as a phenomenal state of a certain type—a state, perhaps, that has a certain intensity, bodily location, and relations to other experiences. Categorical access is a more sophisticated form of access than discriminative access. Just as it is possible to discriminate a bird from its surroundings without being able to recognize it as a bird—perhaps all one can do is bring it under the demonstrative, “that thing there in the sky”—so too it may be possible to discriminate a phenomenal state without being able to recognize it as the kind of phenomenal state it is. Mature human beings enjoy some degree of categorical and discriminative access to their phenomenal states, but many conscious creatures—non-linguistic animals and young children, for example—may enjoy only discriminative access to consciousness.⁴

With this in mind, we can distinguish two forms of introspective optimism. Moderate introspective optimism holds that being in a phenomenal state typically brings with it the capacity to discriminate that state from its phenomenal neighbours, while a more radical form of introspective optimism holds that being in a phenomenal state typically brings with it the capacity to both discriminate and accurately categorize it. By the same token, introspective

pessimism can be more or less radical depending on whether its scope is restricted to categorical access (moderate) or includes both categorical and discriminative access (radical). In what follows, I use the terms “introspective optimism” and “introspective pessimism” to refer to the moderate versions of these views unless noted otherwise.

2.1 The phenomenological argument

Although introspective optimism is often assumed rather than explicitly argued for, I think it is possible to discern two lines of argument for it in the literature. Neither argument is conclusive, but taken together they go some way towards justifying the widespread endorsement of introspective optimism.

The first argument is phenomenological: introspection seems to reveal itself as providing a trustworthy source of information about consciousness. In other words, the epistemic security of introspection seems to be something that is manifest in its very phenomenology. Consider [Brie Gertler](#)’s description of what it is like to attend to the experience that is generated by pinching oneself:

When I try this, I find it nearly impossible to doubt that my experience has a certain phenomenal quality—the phenomenal quality it epistemically seems to me to have, when I focus my attention on the experience. Since this is so difficult to doubt, my grasp of the phenomenal property seems not to derive from background assumptions that I could suspend: e.g., that the experience is caused by an act of pinching. It seems to derive entirely from the experience itself. If that is correct, my judgment registering the relevant aspect of how things epistemically seem to me (this phenomenal property is instantiated) is directly tied to the phenomenal reality that is its truthmaker. (2012, p. 111)

I suspect that Gertler’s comments will strike a chord with many readers—they certainly resonate with me. Introspection seems not merely to

⁴ This claim would need to be tempered if as seems plausible discriminative access requires a minimal form of categorical access. Consider again the case of discriminating a bird but failing to recognize it as a bird. This counts as a failure of categorical access insofar as one fails to bring it under the concept <bird> (or related concepts such as <robin>), but it is arguable that in order to discriminate it from its perceptual background one (or one’s visual system) must bind the various visual features together as the features of a single object, which may require a minimal form of categorical access to the object.

provide one with information about one's experiences, it seems also to "say" something about the quality of that information. This point can be illuminated by contrasting introspection with other forms of access to consciousness. Suppose that you believe that you have the phenomenology associated with anger because a friend has pointed out that you are behaving angrily. In cases like this, testimony provides one with a form of access to one's phenomenal states, but this access surely lacks the epistemic security that introspective access typically possesses—or at least seems to possess. It would be very odd to put more faith in "third-person" evidence concerning one's own conscious states than "first-person" evidence.

Now, one might think that even if the phenomenological consideration just surveyed can *explain* why optimism seems so compelling, it surely can't provide any *justification* for it. Appealing to introspection itself in order to establish its epistemic credentials would be as futile as attempting to pull oneself up by one's own shoelaces. If it's introspection itself that is in the dock, how could its own testimony exonerate it?

In considering this objection we need to distinguish two questions. One question is whether introspection makes claims about its own veracity. A second question is what to make of such claims should they exist—that is, whether to regard them as providing additional reasons for thinking that introspection is trustworthy. Beginning with the first question, it seems to me not implausible to suppose that introspection *could* bear witness to its own epistemic credentials. After all, perceptual experience often contains clues about its epistemic status. Vision doesn't just provide information about the objects and properties present in our immediate environment, it also contains information about the robustness of that information. Sometimes vision presents its take on the world as having only low-grade quality, as when objects are seen as blurry and indistinct or as surrounded by haze and fog. At other times visual experience represents itself as a highly trustworthy source of information about the world, such as when one takes oneself to have a clear

and unobstructed view of the objects before one. In short, it seems not implausible to suppose that vision—and perceptual experience more generally—often contains clues about its own evidential value. As far as I can see there is no reason to dismiss the possibility that what holds of visual experience might also hold true of introspection: acts of introspection might contain within themselves information about the degree to which their content ought to be trusted.

The foregoing addresses the first of the two questions identified above but not the second, for nothing in what I have said provides any reason to think that introspection is a reliable witness to its own veracity. It is one thing for introspection to represent its deliverances *as* trustworthy but it is another for those deliverances to *be* trustworthy. But this being noted, it seems to me not unreasonable to think that the claims introspection makes on its own behalf should be afforded *some* degree of warrant. In general, we regard perceptual testimony as innocent unless proven guilty, and even if introspection is not itself a form of perception it seems reasonable to apply that same rule here. (After all, it is not clear why we would have acquired a cognitive capacity if its deployment routinely led us astray.) The phenomenological argument certainly doesn't provide any kind of proof for introspective optimism, but it seems to me to do more than merely explain why optimism is so attractive: it also provides it with some degree of justification.

2.2 The conceptual argument

A rather different argument for optimism takes as its point of origin the very notion of a phenomenal state. By definition, a phenomenal state is a state that there is "something that it's like" for the subject in question to be in. Conscious creatures enjoy mental states of many kinds, but it is only phenomenal states that bring with them a subjective perspective. But—so the argument runs—if a phenomenal state is a state that there is something it is like to be in, then the subject of that state *must* have epistemic access to its phenomenal character. A

state to which the subject had no epistemic access could not make a constitutive contribution to what it was like for that subject to be the subject that it was, and thus it could not qualify as a phenomenal state. Call this the *conceptual argument*.⁵

How compelling is this argument? It seems to me that a lot depends on what is implied by the notion of “epistemic access”. There is little to recommend the conceptual argument if “epistemic access” is understood in terms of categorization, for it seems fairly clear that a subject need not possess the capacity to accurately categorize its phenomenal states in order for them to contribute to its phenomenal perspective. Of necessity any phenomenal state will fall under categories of various kinds, but the nature of these categories need not be transparent to the creature experiencing it.

But suppose that we construe epistemic access in terms of categorization, rather than identification. Might the conceptual argument justify a moderate form of optimism, according to which subjects must have discriminative access to their phenomenal states? To make this clearer, suppose that it is possible for phenomenal states to occur within the modules of early vision of the kind that are concerned with determining (say) texture or colour constancy. Such phenomenal states—assuming that they are possible—would be completely inaccessible to the subject in question. The creature in question would be unable to contrast the phenomenal character of these states with the phenomenal character of any of its other experiences; it would be unable to single such states out for attention, and it would be unable to make them the objects of demonstrative thought. As such, it seems to me that it is very plausible to hold that they couldn’t be genuinely ascribed to the subject in question, but could at best be ascribed only to one of the subject’s perceptual modules. The root of this intuition, I suspect, lies with the thought that a phenomenal state to which the subject has no discriminative access couldn’t be anything “to” the subject—

that in the relevant sense of the phrase there couldn’t be anything “that it’s like” for the subject to have the relevant experiences.

Although attractive, this argument is not without its problems. One challenge comes in the form of creatures that lack introspective capacities. A creature without introspective capacities might be able to use its conscious states to discriminate some features of the world from others, but it would not be able to make its conscious states themselves objects of its own discriminative activities. And yet—the objection runs—it would be implausible to hold that creatures that lack the capacity for introspective discrimination cannot have phenomenal states. Intuitively, having phenomenal states is one thing and being able to discriminate one’s phenomenal states for each other is another—and more sophisticated—thing. Thus—the argument runs—discriminative access to a phenomenal state cannot be a necessary condition for being in that state.

I certainly agree that it would be implausible to restrict phenomenal states to creatures that possess introspective capacities, but perhaps the objection can be met without making such a restriction. What we can say is that when a creature does acquire introspective capacities those capacities bring with them the ability to discriminate its phenomenal states from one another (at least under epistemically benign conditions). So, we can grant that being in a phenomenal state doesn’t require discriminative access to that state, but also hold that creatures with introspective capacities will be able to discriminate their phenomenal states from one another (again, at least when conditions are epistemically benign).

A second objection to the conceptual argument concerns states that occupy the “margins” of consciousness—such as the unnoticed hum of the refrigerator or the background phenomenology of mood experiences. It is arguable that in some cases experiences like this not only fail to fall within the scope of introspection but in fact cannot be brought within its scope, for to attend to them would be to bring them into the “centre” of consciousness and thus change their phenomenal character. Such states serve as po-

⁵ There are echoes here of the claim that phenomenal consciousness entails a certain kind of “access consciousness”. For some relevant discussion see Church (1997) and Clark (2000).

tential counter-examples to the claim that creatures with introspective capacities must be able to discriminate their phenomenal states from one another.

In response, one might grant that even if the phenomenal states that occur in the margins of consciousness cannot be singled out for introspective attention, there is still a sense in which they can be the objects of discrimination. Not only can they be discriminated from one another, they can also be discriminated from those phenomenal states that *do* fall within the scope of attention. Indeed, if such states cannot be discriminated from their phenomenal neighbours in any way then it is unclear what reason we could have for thinking of them as falling within the margins of *consciousness* at all, rather than being completely unconscious.

Where do these considerations leave us? I have suggested that the phenomenological argument provides some reason to take at least a moderate form of optimism seriously. It doesn't, of course, establish that our access to all kinds of phenomenal states is robust—indeed, one might even appeal to phenomenological considerations to motivate the idea that our epistemic access to significant regions of phenomenal space is very poor. (I return to this topic shortly.) The conceptual argument provides little reason to think that we will always be able to *categorize* our phenomenal states, but it does provide some motivation for the idea that being in a phenomenal state brings with it the ability to *discriminate* that phenomenal state, at least when it comes to creatures with introspective capacities. In short, optimism of at least a moderate form is not merely a holdover from Cartesianism but can be provided with some degree of support. With these considerations in mind let us turn now to the case for pessimism.

3 Motivating pessimism

Two distinctions will prove helpful in what follows. One distinction is between forms of pessimism that concern only our capacity to identify our phenomenal states and forms of pessimism that call into question our capacity to both discriminate and categorize our phenomenal states. A

second distinction concerns the *scope* of pessimism. At one end of the spectrum are local forms of pessimism that concern only a relatively circumscribed range of phenomenal states (say, imagery experiences), while at the other end of the spectrum are forms of pessimism that are unrestricted in scope. Perhaps no theorist has ever embraced a truly global form of pessimism—even Schwitzgebel grants that introspection is trustworthy with respect to certain aspects of consciousness—but some forms of pessimism are clearly wider in scope than others. These two distinctions are, of course, orthogonal to each other. One could be a moderate but global pessimist; alternatively, one could endorse a radical but highly local form of pessimism.

So much for the varieties of pessimism—how might one argue for the view? One influential line of argument for pessimism—or at least something very much like it—appeals to the alleged privacy of introspection. Because an individual's introspective judgments cannot be checked by anyone else, it follows—so the argument runs—that it would be inappropriate to trust them. This argument is often used to motivate the view that introspection is scientifically illegitimate, but it could also be used to motivate the view that one should adopt a sceptical attitude towards one's own introspective capacities.⁶ Although it has been influential, I will leave this argument to one side in order to focus on a trio of arguments that aim to establish not merely that there is no positive reason to trust introspection (as the argument just outlined attempts to do), but that there is positive reason *not* to trust it. My presentation of these arguments will draw heavily on Schwitzgebel's work, for he has done more than any other author to develop and defend them.⁷

But before I examine those arguments, I want to consider the overall structure of

6 For critical discussion of this argument see Goldman (1997, 2004) and Piccinini (2003, 2011). In my view the most plausible response to it involves denying that introspection is private in the sense required for the argument to go through. I touch briefly on this idea in section 4.

7 Schwitzgebel is clearly attracted to a fairly global form of introspective pessimism, but (to the best of my knowledge) he doesn't distinguish between discriminative and categorical access, and thus it is unclear whether his version of pessimism is radical or merely moderate. Generally, however, he seems to have something akin to radical scepticism in mind.

Schwitzgebel's case for global scepticism. As I read him, Schwitzgebel employs a two-step strategy (2008, p. 259). The first step involves attempting to establish a form of local pessimism via one (or more) of the three argumentative strategies to be explored below. The second step involves generalizing from the kinds of phenomenal states that are the targets of local pessimism to phenomenal states in general. The second step is clearly required, for without it we would have no reason to regard introspection *in general* as “faulty, untrustworthy, and misleading”—“not just possibly mistaken, but massively and pervasively” (Schwitzgebel 2008, p. 259).⁸

I will consider both steps in due course, but the crucial point to note for now is that, considered in the abstract, the second step of the argument looks somewhat suspect (Bayne & Spener 2010). Even if there are hard cases for introspection—that is, cases in which introspective access to phenomenology is insecure—there also easy cases—that is, cases in which introspective access to phenomenology is clearly secure. Indeed, Schwitzgebel himself grants that introspection “may admit obvious cases” and that some aspects of visual experience “are so obvious it would be difficult to go wrong about them” (Schwitzgebel 2008, p. 253). But if that's the case, then one might well ask why we shouldn't generalize from those cases rather than from the hard cases on which he focuses. Schwitzgebel complains that to generalize about introspection only on the basis of the easy cases “rigs the game”. That's true. But it's equally true that to generalize only on the basis of the hard cases—as Schwitzgebel seems to do—would *also* rig the game. In fact, it would seem pretty clear that any comprehensive account of the epistemic landscape of introspection must

take *both* the hard and easy cases into consideration. Arguably, generalizing beyond the obviously easy and hard cases requires an account of what makes the hard cases hard and the easy cases easy. Only once we've made some progress with that question will we be in a position to make warranted claims about introspective access to consciousness in general. What this suggests is that although there is a formal distinction between the two steps of Schwitzgebel's argument, the steps are not entirely independent of each other, for the fortunes of the second step rest in part on the case that can be made for the first step. With that thought in mind, let us now turn to the arguments for pessimism.

3.1 The argument from dumbfounding

One line of argument that features prominently in Schwitzgebel's work is what I call the argument from dumbfounding.⁹ Arguments of this form involve posing introspective questions that allegedly stump us—questions that we find ourselves unable to answer with any significant degree of confidence. Here's an example of such an argument:

Reflect on, introspect, your own ongoing emotional experience at this instant. Do you even have any? If you're in doubt, vividly recall some event that still riles you until you're sure enough that you're suffering from renewed emotion. Or maybe your boredom, anxiety, irritation, or whatever in reading this essay is enough. Now let me ask: Is it completely obvious to you what the character of that experience is? Does introspection reveal it to you as clearly as visual observation reveals the presence of the text before your eyes? Can you discern its gross and fine features through introspection as easily and as confidently as you can, through vision, discern the gross and fine features of nearby external objects? Can you trace its spatiality (or non-spatiality), its viscosity or cognitiveness, its involvement with conscious

⁸ Another reconstruction of Schwitzgebel's overarching argumentative strategy proceeds as follows. Although the arguments from dumbfounding, dissociation, and variation establish only local forms of introspective pessimism when considered on their own, when taken collectively they provide a good case for a relatively global form of pessimism given that each of the three arguments concerns distinct (albeit, perhaps, overlapping) domains of phenomenology. Thus understood, Schwitzgebel does not need to appeal to a generalization from the “hard cases” to introspection in general. Although this construal provides an alternative route to pessimism, I regard it as less promising than the one outlined in the text—both as a reading of Schwitzgebel's work and as an argument in its own right.

⁹ Following Hohwy (2011), Schwitzgebel (2011) calls this “the argument from uncertainty”.

imagery, thought, proprioception, or whatever, as sharply and infallibly as you can discern the shape, texture and color of your desk? (Or the difference between 3 and 27?) I cannot, of course, force a particular answer to these questions. I can only invite you to share my intuitive sense of uncertainty. (Schwitzgebel 2008, p. 251)

This argument does not appeal to independent evidence in order to motivate pessimism. Rather, it appeals to first-person considerations: introspection *itself* seems to suggest that there are aspects of our own conscious experience that elude our grasp. As Schwitzgebel puts it, “it’s not just language that fails us—most of us?—when we confront such questions [...] but introspection itself. [...] in the case of emotion, the very phenomenology itself—the qualitative character of our consciousness—is not entirely evident” (Schwitzgebel 2008, pp. 249–250).

Before examining the force of this argument, let us first consider what kind of pessimism it aims to establish. Does the above passage call into question our capacity to accurately *categorize* our emotional phenomenology, or is the claim rather that we lack even the capacity to *discriminate* our emotional experiences from one another and from the rest of our phenomenal states? Although Schwitzgebel’s concern seems to include questions of discriminative access—after all, the passage begins by asking if we can even tell whether or not we have any emotional phenomenology—I take his worries to centre on our capacity to accurately categorize our emotional phenomenology. As I read him, Schwitzgebel’s questions focus on our ability to determine how our emotional experience is structured, both internally and in terms of its relations to phenomenal states of other kinds.

I think that the questions Schwitzgebel raises *are* difficult to answer. However, it is not clear to me that this fact provides quite as much support for introspective pessimism as Schwitzgebel thinks it does. Lying behind the dumbfounding strategy is the assumption that the questions being posed have determinate answers—that they are appropriate questions to ask. However, I suspect that in an important

range of cases this assumption may be unjustified. With respect to the phenomenology of emotion it is natural to assume that the boundaries between the phenomenal states associated with emotion are as clean and sharp as the boundaries between our standard ways of categorizing emotional states. We regard boredom, anxiety, and irritation as distinct emotional states, and we also regard each of these states as associated with distinctive forms of phenomenology. On the basis of these two thoughts we assume that the phenomenal states associated with these categories can themselves be cleanly distinguished from one another. Thus, when one finds oneself at a loss to know whether one is in *the* phenomenal state associated with boredom, anxiety, or irritation one naturally assumes that the fault lies with one’s introspective capacities. But perhaps the mistake was to assume that the phenomenology of emotion can be cleanly demarcated into states that are uniquely associated with either boredom, anxiety, or irritation. Perhaps the phenomenal states associated with these emotional states overlap and interpenetrate each other. If this were the case, then although there might be certain contexts in which one’s emotional phenomenology is purely that of (say) boredom, there may also be other contexts in which one’s emotional phenomenology involves a complex mix of the phenomenal states associated with boredom, anxiety and irritation. And if one were in a context like this, one might be at something of a loss to know just how to categorize one’s emotional state. The only categories that might come to mind would be those associated with the folk psychology of emotion—*<boredom>*, *<anger>* and *<irritation>*—but these categories might fail to cut the phenomenology of emotion at its joints. In other words, emotional phenomenology may pose a particular introspective challenge not because introspection does a poor job of acquainting us with emotional phenomenology, but because the structure of the phenomenology of emotion fails to map onto the structure of our folk categories of emotions in a straightforward manner.

Other versions of the argument from dumbfounding raise a different set of challenges

for introspective optimism. Consider the question of introspective access to visual imagery. Schwitzgebel asks his readers to form a visual image of the front of his or her house, and to then consider the following questions:

How much of the scene are you able vividly to visualize at once? Can you keep the image of your chimney vividly in mind at the same time you vividly imagine (or “image”) your front door? Or does the image of your chimney fade as your attention shifts to the door? If there is a focal part of your image, how much detail does it have? How stable is it? Suppose that you are not able to image the entire front of your house with equal clarity at once, does your image gradually fade away towards the periphery, or does it do so abruptly? Is there any imagery at all outside the immediate region of focus? If the image fades gradually away toward the periphery, does one lose colours before shapes? Do the peripheral elements of the image have color at all before you think to assign color to them? Do any parts of the image? If some parts of the image have indeterminate colour before a colour is assigned, how is that indeterminacy experienced—as grey?—or is it not experienced at all? If images fade from the centre and it is not a matter of the color fading, what exactly are the half-faded images like? (Schwitzgebel 2002, pp. 38–39)

I think that this line of questioning poses one of the most significant challenges to optimism. Further, it is doubtful whether this challenge can be resisted in the way that the previous version of the dumbfounding challenge can, for these questions don’t seem to rest on any problematic assumptions. Schwitzgebel isn’t assuming that visual imagery must be pictorial in nature, or that it will always be fully detailed and determinate. Rather, one issue that he explicitly puts on the table is whether the phenomenology of visual imagery can be purely “generic” or “gisty”, or whether it must instead always be specific in some way or another.

But perhaps the dumbfounding challenge can be met in another way. As Jakob Hohwy (2011) has noted, one striking feature of visual imagery is its instability:

In the absence of specific goal parameters for simulations there will be much phenomenal variability because in such conditions subjects must themselves make up the purposes for which they imagine things, or engage in ‘simple’ free-wheeling imagery. For example, there is an indefinite number of purposes for which you can imagine the front of your house (walking up to it, standing close by, assessing its shape, its prettiness, flying around it, how the postman sees it, smelling it, repairing it, buying it, selling it etc), each of these purposes will constrain the imagery, and thus the introspected phenomenology, in different ways. This means that subjects probably do have *variable* phenomenology, and introspectively report so reliably. (2011, p. 279)

Hohwy’s comments are intended to explain the variability in the introspective reports that individuals give, but they also bear on the dumbfounding argument. Perhaps we are not sure how best to describe the phenomenology of imagery because it is so variable. Imagery experiences cannot be pinned down, but are constantly shifting in response to our own imagistic activity. Precisely how much of the scene we vividly visualize “all at once” depends on the goals that constrain the act of visualization. And, as Hohwy suggests, when we have no such goals our imagery may end up “freewheeling”, such that we move from one state to another. Hohwy grounds his analysis in a predictive-coding account of cognition, but his fundamental point is independent of that theoretical framework and should be fairly uncontentious: imagery surely *is* more labile than perceptual experience or bodily sensation. No wonder, then, that its phenomenal structure is that much more difficult to articulate.

I have suggested that the optimist has the resources to meet (or at least “problematize”)

two of the leading versions of the dumbfounding argument. But suppose that my responses are found wanting, and that the pessimist is able to show that our introspective access to both emotional and imagery phenomenology is insecure and impoverished. Even so, there would be a further question as to how such a finding would motivate *global* pessimism. It is certainly true that questions about the nature of certain kinds of experiences (e.g., emotional and imagery experiences) strike us as difficult to answer and may leave us flummoxed, but it is equally true that many introspective questions strike us as easily answered. Indeed, as the quotation from Gertler makes vivid, many of our introspective judgments appear to be accompanied by a sense of epistemic certainty. Why should we generalize from the first set of cases rather than the second? Without an account of why certain introspective questions leave us dumbfounded it is difficult to see why pessimism about a particular range of introspective questions should undermine the epistemic credentials of introspection more generally. So even if the threat posed by dumbfounding arguments were able to establish a form of local pessimism, that threat would appear to be easily quarantined.

3.2 Dissociation arguments

A very different case for introspective pessimism is provided by what I call dissociation arguments. Such arguments appeal to a lack of congruence between a subject's introspective judgments and their capacity to produce reliable first-order judgments—that is, judgments about the objects and properties in their environment. An example of this kind of argument is provided by Schwitzgebel's treatment of the so-called “grand illusion” (Noë 2002). Most people, Schwitzgebel claims, hold that a broad swathe of their environment—perhaps thirty or more degrees—is clearly presented within visual experience with its “shapes, colours, textures all sharply defined”. Schwitzgebel argues that we have good reason to regard such claims as false. In making the case for this claim, he appeals to an example first popularized by Dennett (1991):

Draw a card from a normal deck without looking at it. Keeping your eyes fixed on some point in front of you, hold the card at arm's length just beyond your field of view. Without moving your eyes, slowly rotate the card toward the centre of your visual field. How close to the centre must you bring it before you can determine the colour of the card, its suit, and its value? Most people are quite surprised at the result of this little experiment. They substantially overestimate their visual acuity outside the central, foveal region. When they can't make out whether it's a Jack or a Queen though the card is nearly (but only nearly) dead centre, they laugh, they're astounded, dismayed. (Schwitzgebel 2008, pp. 254–255)

How might we explain the dissociation between subjects' introspective judgments and their first-order judgments? One explanation is that the subjects' introspective beliefs are false, and that people wrongly take themselves to have detailed visual phenomenology outside of the focus of attention. This is the explanation that Schwitzgebel endorses. But as Schwitzgebel (2008, p. 255) himself notes, it is possible to explain this dissociation by supposing that individuals are wrong not about which phenomenal states they are in but only about the *origin* of that state. With respect to the card trick example, the proposal is that subjects do indeed have detailed visual phenomenology outside of the origin of attention, but that this phenomenology derives from background expectation rather than environmental input—that is, it is “illusory”.

Schwitzgebel's account of the dissociation may have more intuitive appeal than the account I have just outlined, but it is not clear how the data furnished by the dissociation argument allows us to choose between them. However, reasons to favour Schwitzgebel's account can be gleaned noting that the judgment on which we have focused—“thirty or more degrees of my visual field presents itself to me clearly in experience with its shapes, colours, textures all sharply defined”—is available to

introspection only indirectly. This judgment is not the direct reflection of any one introspective act, but is a belief about the nature of one's visual experience that one forms by tracking one's introspective capacities over time. Call such judgments *indirectly introspective*. Indirectly introspective judgments can be contrasted with *directly introspective* judgments—that is, judgments of the kind that one makes in the very context of the card trick experiment, such as “I am now experiencing the shape, colour, and texture of this card (which is presented to me slightly off centre) in sharp detail”. We can now see that although there is a dissociation between the first-order judgments that subjects make and their *indirect* introspective judgments, there is no such dissociation between their first-order judgments and their *direct* introspective judgments. Subjects in the card-trick experiment *don't* report experiencing the shape, colour, and texture of cards that are presented slightly off centre to them “in sharp detail”—rather, they claim to lack sharp and detailed experiences of such objects. Direct introspective judgments clearly have more warrant than indirect judgments, and thus there is good reason to prefer Schwitzgebel's explanation of the dissociation over the alternative account.

But although we have found reasons to support Schwitzgebel's analysis of the dissociation, we have seen that these very reasons undermine his pessimistic attitude to introspection in general, for the evidence in favour of Schwitzgebel's account involves an appeal to introspection. In other words, the pressure that the dissociation argument puts on indirect introspective judgments assumes that direct introspective judgments are trustworthy. The card trick case does indeed cast doubt on the epistemic security of our *background* beliefs about our own visual experience, but there is no reason to extend such doubts to include our direct introspective judgments; and it is surely direct introspective judgments that are at the heart of debates about the trustworthiness of introspection. (Indeed, indirect introspection judgments are not really a genuine form of introspection at all.)

Let us turn now to the second step of the dissociation argument: the inference from local pessimism to general pessimism. Suppose that we were to find a dissociation between a certain range of introspective judgments and the subject's capacity to make the corresponding first-order judgments. Suppose, furthermore, that one could show that this dissociation is best explained by assuming that the introspective judgments in question were false. Would one have any reason to think that introspection *in general* ought to be regarded with suspicion? Not as far as I can see. It seems to me that our faith in the robustness of introspective access to domains in which such dissociations are not to be found ought to remain completely untroubled by such a finding. In fact, one might even argue that coherence between first-order judgments and (direct) introspective judgments would provide evidence in favour of introspective optimism. If dissociations between a person's introspective capacities and their first-order capacities can *disconfirm* their introspective judgments (as the dissociation argument assumes), then *associations* between a person's introspective judgments and their first-order capacities ought to *confirm* them (Bayne & Spener 2010). In other words, the fact that a person's introspective judgments cohere with their capacity to produce reliable reports of their environment ought to provide us with positive reason to trust those judgments.¹⁰ And a great number of our introspective reports clearly *do* cohere with our first-order capacities. Although there are cases in which such coherence fails to obtain—for example, Schwitzgebel (2011, Ch. 3) provides a plausible case for the claim that introspective reports of visual imagery are only weakly correlated with the kinds of first-order cognitive capacities that one would expect visual imagery to subserve—such cases

¹⁰ This argument is closely related to an argument presented by Spener (2013) in defence of the idea that we can provide principled reasons for trusting introspection in certain contexts. Spener argues that certain everyday abilities, such as adjusting a pair of binoculars or ordering food in a restaurant, are introspection-reliant—that is, their successful execution requires that the subject have accurate introspective judgments. I find Spener's argument plausible, but, as Schwitzgebel (2013) notes, it is something of an open question just how many of our everyday abilities are reliant on introspection. At any rate, the argument I have given here makes no appeal to that notion.

are striking precisely because they stand out against the backdrop of coherence that characterizes the relationship between our normal introspective reports and our first-order perceptual capacities.

3.3 Arguments from introspective variation

Perhaps the strongest case for introspective pessimism derives from the phenomenon of introspective variation. Such arguments have as their starting point a disagreement about how best to describe some aspect of phenomenology. Pessimists then argue that the best explanation for the introspective dispute is that at least one of the two groups is mistaken about its own phenomenology, and thus that introspective access to the relevant phenomenal domain is insecure: despite their best efforts, at least one of the two parties to the dispute is wrong about its own phenomenology.

Schwitzgebel (2008) examines a number of arguments from introspective variation, but his central case study concerns a debate about the nature of conscious thought—the so-called “cognitive phenomenology” debate (Bayne & Montague 2011; Smithies 2013).¹¹ On one side of this dispute are those who deny that thought has a distinctive phenomenal character. Those who hold this view typically allow that conscious thought has a phenomenology of some kind, but they regard that phenomenology as purely sensory—as limited to the phenomenology of inner speech, visual imagery, and so on. We might call this the conservative account of conscious thought, for it treats phenomenal consciousness as limited to sensory aspects of the mind. On the other side of this dispute are those who adopt a liberal conception of conscious thought, according to which conscious thought is characterized by a range of non-sensory phenomenal states—states of “cognitive phenomenology”. It is tempting to conclude that at least one of these two sides is guilty of a

fairly radical introspective error: introspection either fails to inform conservatives of a wide range of phenomenal states that they enjoy on a regular basis, or it misleads liberals into thinking that they enjoy a wide range of phenomenal states that they don’t enjoy. Either way, introspection would seem to be untrustworthy with respect to what is clearly a central feature of phenomenology.¹²

But before we follow Schwitzgebel (and many others) in embracing this conclusion, we need to consider alternative explanations of the cognitive phenomenology dispute. One possible explanation appeals to group differences in phenomenology. Perhaps the descriptions of conscious thought that both liberals and conservatives give are right when applied to themselves but wrong when taken to describe conscious thought in general. In other words, perhaps both parties to the dispute are guilty of over-over-hasty generalization rather than introspective error.

Although an appeal to group differences might explain (away) some instances of introspective disagreement, it is unlikely to provide the best explanation of the cognitive phenomenology dispute. First, this account requires a degree of variation in phenomenology for which there are few (if any) parallels. This is not to say that phenomenal differences between individuals might not run much deeper than common-sense tends to assume—consider, for example, the phenomenal differences that characterize synaesthesia (Robertson & Sagiv 2005)—but the kinds of phenomenal differences that we already recognize are nowhere near as fundamental as the kinds of differences required by this explanation of the cognitive phenomenology debate, for liberals claim that conscious thought is characterized by a *sui generis* kind of phenomenology—a kind that is non-sensory in nature. Second, the group difference proposal predicts that there are cognitive and behavioural differences between the advocates of cognitive phenomenology and their detractors that simply don’t appear to obtain. In sum, it seems

¹¹ Other examples of recent introspective disagreement concern the apparent shape of the objects of visual experience (e.g., Siewert 2007; Schwitzgebel 2011, Ch. 2), the existence of high-level perceptual phenomenology (Siegel 2006; Bayne 2009), and the satisfaction conditions of the phenomenology of free will (e.g., Horgan 2012; Nahmias et al. 2004).

¹² The conservative view is also known as the “restrictive” (Prinz 2011) or “exclusivist” (Siewert 2011) view, while the liberal view is also known as the “expansionist” (Prinz 2011) or “inclusivist” (Siewert 2011) view.

highly unlikely that the debate about the existence of cognitive phenomenology can be explained by supposing that what it is like to be a liberal is different from what it is like to be a conservative.

But there is another deflationary explanation of the debate about cognitive phenomenology that cannot be so easily dismissed. Perhaps the parties to the debate are operating with very different conceptions of what it would take for thought to possess distinctive phenomenal character, and are thus talking passed each other (Bayne unpublished). On this proposal, liberals are willing to extend the notion of phenomenal consciousness beyond its sensory paradigms in a way that conservatives are not. If this account is right, then the dispute surrounding the existence of cognitive phenomenology is largely verbal. Rather than disagreeing about what introspection reveals, the two sides instead disagree about how the term “phenomenal consciousness” and its cognates ought to be employed.

Why take this proposal seriously? Well, one argument for it is that it would provide a good explanation of why there is such widespread disagreement about the nature of conscious thought—the very terms in which the debate are couched are contested. It is also widely acknowledged that there are different notions of “what it’s likeness” (see e.g., Tye 1996; Flanagan 1992; Georgalis 2005). Although this proposal clearly needs much more defence and development than I can give it here, I think it is not unreasonable to suppose that the disagreement surrounding the existence of cognitive phenomenology might turn out to be largely verbal. At any rate, it seems to me that this account provides at least as good an explanation of the dispute as that which is required by the argument from variation.¹³

¹³ Of course, the pessimist might argue that, even if the disagreement surrounding the phenomenology of thought is fundamentally semantic, it doesn’t follow that the optimist is off the hook. After all, using introspection to ground a science of consciousness doesn’t merely require the reliability of introspection, it also requires intersubjective agreement about its deliverances. And—the pessimist might continue—dispute about how to apply the term “phenomenal consciousness” and its cognates threatens to undermine intersubjective disagreement about what introspection reveals just as surely as introspective unreliability does. This is a

There are, of course, other introspective disagreements besides that concerning the phenomenology of thought, and nothing that I have said here goes any way towards showing that they too succumb to a deflationary analysis. Indeed, I suspect that certain introspective disputes—for example, those relating to the richness of visual imagery—may well be best explained by appeal to introspective error. But even if the argument from variation succeeds in establishing a local form of pessimism, it seems to me there is little reason to think that this pessimism generalizes. Indeed, domains that feature disagreement in introspective reports stand out against a general backdrop of introspective agreement. Arguably many domains of consciousness exhibit a great deal of uniformity with respect to introspective reports once individual differences and verbal disputes are taken into account. Now, although inter-subjective agreement doesn’t entail that the individuals in question are right, it does need to be explained, and it seems plausible to suppose that leading explanations of inter-subjective agreement will appeal to the trustworthiness of introspection.

4 Elusive phenomenology

In the previous section I argued that there are good reasons for resisting Schwitzgebel’s case for global pessimism. However, we also saw that there are domains in which our introspective access to phenomenal consciousness is rather less secure than we might have pre-theoretically assumed. In other words, we saw that there is reason to think that certain kinds of phenomenal states are introspectively elusive. In this final section I want to sketch an account of why certain types of phenomenal states are elusive and others are not.

Let me begin by distinguishing the form of phenomenal elusiveness with which I am concerned from another notion of phenomenal elusiveness that I want to set to one side. In a recent paper, Kriegel uses the label “elusive phe-

fair challenge, but in my view the prospects for securing a solution to the cognitive phenomenology dispute, should it turn out to be fundamentally semantic, are quite high. For further discussion of phenomenal disputes and introspective disagreement see Hohwy (2011) and Siewert (2007).

nomenology” to describe phenomenal states “whose very essence requires the absence of introspective attention” (2013, p. 1171). Among the examples that he gives of elusive phenomenology are the phenomenal states that occur at the fringes or margins of consciousness. As Kriegel notes, such states are elusive in that any attempt to make them the object of attentive introspection would change their nature. Although Kriegel’s notion of elusiveness is closely related to the one that I employ here, the two notions are not identical. (One way of seeing that they are distinct is that Kriegel’s elusiveness is primarily a matter of the phenomenology, whereas my elusiveness is a matter of one’s introspective access to the phenomenology.) Unlike Kriegel, I am interested in a type of elusiveness that is independent of attention. Consider again visual imagery. Although particular instances of visual imagery might be elusive in Kriegel’s sense because they happen to occupy the margins of consciousness, I am interested here in the fact that visual imagery *as such* appears to be introspectively elusive.¹⁴

Why might certain types of phenomenal states be elusive in a way that other types of phenomenal states are not? Broadly speaking, there are two places in which we might look for an answer to this question. On the one hand we might appeal to intrinsic features of the phenomenal states themselves. Perhaps there is something inherent in the very nature of certain kinds of phenomenal states that renders them relatively opaque to introspective access. Another possibility is that the elusiveness of certain types of phenomenal states has nothing to do with their intrinsic nature but instead reflects the structure of our introspective capacities. Just as our perceptual system is geared toward the identification of certain kinds of environmental states rather than others, so too it is possible that our introspective system is geared towards the identification of certain kinds of phenomenal states rather than others. On this view, the fact that our introspective access to some types of phenomenology is more secure

than it is to others tells us more about introspection than it tells us about phenomenal consciousness (as it were).

It is, I think, premature to speculate which of these two accounts might be the more plausible; indeed, it is possible that a full explanation of elusiveness will have to draw on both ideas. But rather than pursue that thought, I want instead to sketch one way in which the structural features of introspection might go some way towards explaining why certain types of introspective judgments are more secure than others. The account in question appeals to a distinction between two kinds of introspective judgments: *scaffolded judgments* and *freestanding judgments* (Bayne & Spener 2010). The distinction is perhaps best grasped by means of examples. Contrast an introspective judgment that is directed towards one’s visual experience of looking at a red tomato with an introspective judgment that is directed towards an experience of visual imagery involving a red tomato in front of one. In the former case, there is a perceptual judgment that one is disposed to make (“There is a red tomato in front of me”) whose content corresponds (broadly speaking) to the content of one’s introspective judgment (“I have an experience as of a red tomato in front of me”). In the latter case, however, there is no such first-order judgment that one is disposed to make whose content might correspond to the content of one’s introspective judgment. In a sense, the former judgment is “scaffolded” by a perceptual disposition in a way that the latter judgment is not.

I suggest that scaffolded judgments are typically more secure than freestanding ones precisely because they are scaffolded. At the very least, it is a striking fact that many of the most epistemically insecure introspective judgments appear to be freestanding. Further, one can tell an attractive story about *why* introspective scaffolding might contribute to epistemic security. In making scaffolded judgments, the subject is able to both exploit the resources that it has for making freestanding judgments and calibrate those resources by drawing on its dispositions to make first-order

¹⁴ Phenomenal domains that are at least somewhat elusive include the phenomenology of agency (Metzinger 2006; Bayne 2008; Horgan et al. 2006) and high-level perceptual phenomenology (Siegel 2006; Bayne 2009).

perceptual judgments.¹⁵ Just as beliefs that are derived from multiple (independent) sources are typically more secure than beliefs derived from just a single source, so too scaffolded introspective judgments might typically be more secure than their freestanding brethren.

5 Conclusion

This paper provides a partial response to Schwitzgebel's case for global pessimism with respect to introspection. I began by outlining two arguments for optimism; the first argument turned on an appeal to the phenomenology of introspection, while the second drew on a conceptual connection between the notions of introspective access and phenomenality. Neither argument comes close to being decisive, but taken together they provide some explanation for—and justification of—the widespread appeal of optimism. I then turned to a detailed examination of Schwitzgebel's case for pessimism, arguing that although his arguments go some way towards justifying local pessimism (particularly with respect to imagery), there is little reason to generalize that pessimistic attitude to introspection more generally.

But perhaps the central lesson of this paper is that the epistemic landscape of introspection is far from flat but contains peaks of security alongside troughs of insecurity. Rather than asking whether or not introspective access to the phenomenal character of consciousness is trustworthy, we should perhaps focus on the task of identifying how secure our introspective access to various kinds of phenomenal states is, and why our access to some kinds of phenomenal states appears to be more secure than our access to other kinds of phenomenal states. I have suggested that the notion of introspective scaffolding might play a role in answering this second question, but that that proposal is at

best only a very small part of a much larger account of introspective insecurity. There is certainly a lot more work to be done before we have a good grip on the epistemic structure of introspection.

Acknowledgements

An earlier version of this paper was presented at a graduate workshop on perception at the City University of New York, and I am grateful to the members of the audience on that occasion for their comments. I also gratefully acknowledge the support of European Research Council Grant *The Architecture of Consciousness* (R115798).

¹⁵ An influential account of introspection holds that introspection involves a semantic ascent routine in which one redeploys rather than represents one's introspective target (Byrne 2005; Evans 1982; Fernández 2013). Although I am not endorsing this account of introspection in general (or indeed of introspective access to perceptual phenomenology in particular), I am suggesting that such procedures might be implicated in introspective access to certain kinds of phenomenal states.

References

- Bayne, T. (2008). The phenomenology of agency. *Philosophy Compass*, 3 (1), 182-202. [10.1111/j.1747-9991.2007.00122.x](https://doi.org/10.1111/j.1747-9991.2007.00122.x)
- (2009). Perception and the reach of phenomenal content. *Philosophical Quarterly*, 59 (236), 385-404. [10.1111/j.1467-9213.2009.631.x](https://doi.org/10.1111/j.1467-9213.2009.631.x)
- Bayne, T. (unpublished). *The puzzle of cognitive phenomenology*.
- Bayne, T. & Montague, M. (2011). Cognitive phenomenology: An introduction. In T. Bayne & M. Montague (Eds.) *Cognitive phenomenology* (pp. 1-34). Oxford, UK: Oxford University.
- Bayne, T. & Spener, M. (2010). Introspective humility. *Philosophical Issues*, 20 (1), 1-22. [10.1111/j.1533-6077.2010.00176.x](https://doi.org/10.1111/j.1533-6077.2010.00176.x)
- Byrne, A. (2005). Introspection. *Philosophical Topics*, 33 (1), 79-104. [10.5840/philtopics20053312](https://doi.org/10.5840/philtopics20053312)
- Chalmers, D. J. (2003). The content and epistemology of phenomenal belief. In Q. Smith & A. Jokic (Eds.) *Consciousness: New philosophical perspectives* (pp. 220-272). Oxford, UK: Oxford University Press.
- Church, J. (1997). Fallacies or analyses? In N. Block, O. Flanagan & G. Güzeldere (Eds.) *The nature of consciousness* (pp. 425-426). Cambridge, MA: MIT Press.
- Clark, A. (2000). A case where access implies qualia? *Analysis*, 60 (1), 30-38. [10.1093/analys/60.1.30](https://doi.org/10.1093/analys/60.1.30)
- Dennett, D. (1991). *Consciousness explained*. Boston, MA: Little, Brown.
- Evans, G. (1982). *The varieties of reference*. Oxford, UK: Oxford University Press.
- Fernández, J. (2013). *Transparent minds: A study of self-knowledge*.
- Flanagan, O. (1992). *Consciousness reconsidered*. Cambridge, MA: MIT Press.
- Georgalis, N. (2005). *The primacy of the subjective*. Cambridge, MA: MIT Press.
- Gertler, B. (2012). Renewed acquaintance. In D. Smithies & D. Stoljar (Eds.) *Introspection and consciousness* (pp. 93-128). Oxford, UK: Oxford University Press.
- Goldman, A. (1997). Science, publicity, and consciousness. *Philosophy of Science*, 64 (4), 525-545. [10.1086/392570](https://doi.org/10.1086/392570)
- (2004). Epistemology and the evidential status of introspective reports. *Journal of Consciousness Studies*, 11 (7-8), 1-16.
- Haybron, D. (2007). Do we know how happy we are? On some limits of affective introspection and recall. *Noûs*, 41 (3), 394-428. [10.1111/j.1468-0068.2007.00653.x](https://doi.org/10.1111/j.1468-0068.2007.00653.x)
- Hohwy, J. (2011). Phenomenal variability and introspective reliability. *Mind & Language*, 26 (3), 261-286. [10.1111/j.1468-0017.2011.01418.x](https://doi.org/10.1111/j.1468-0017.2011.01418.x)
- Horgan, T. (2012). Introspection about phenomenal consciousness: Running the gamut from infallibility to impotence. In D. Smithies & D. Stoljar (Eds.) *Introspection and consciousness* (pp. 405-422). Oxford, UK: Oxford University Press.
- Horgan, T., Tienson, J. & Graham, G. (2006). Inner-world scepticism and the self-presentational nature of phenomenal consciousness. In U. Kriegel & K. Williford (Eds.) *Self-representational approaches to consciousness* (pp. 41-61). Cambridge, MA: MIT Press.
- Horgan, T. & Kriegel, U. (2007). What is phenomenal consciousness that we may know it so well? *Philosophical Issues*, 17 (1), 123-144. [10.1111/j.1533-6077.2007.00126.x](https://doi.org/10.1111/j.1533-6077.2007.00126.x)
- Kriegel, U. (2013). A hesitant defense of introspection. *Philosophical Studies*, 165 (3), 1165-1176. [10.1007/s11098-013-0148-0](https://doi.org/10.1007/s11098-013-0148-0)
- Metzinger, T. (2006). Conscious volition and mental representation: Towards a more fine-grained analysis. In N. Sebanz & W. Prinz (Eds.) *Disorders of volition* (pp. 19-48). Cambridge, MA: MIT Press.
- Nahmias, E., Morris, S., Nadelhoffer, T. & Turner, J. (2004). The phenomenology of free will. *Journal of Consciousness Studies*, 11 (7-8), 162-179.
- Noë, A. (2002). Is the visual world a grand illusion? *Journal of Consciousness Studies*, 9 (5-6), 1-12.
- Piccinini, G. (2003). First-person data, publicity, and self-measurement. *Philosophers' Imprint*, 9 (9), 1-16.
- (2011). Scientific methods must be public, and Descriptive Experience Sampling qualifies. *Journal of Consciousness Studies*, 18 (1), 102-117.
- Prinz, J. (2011). The sensory basis of cognitive phenomenology. In T. Bayne & M. Montague (Eds.) *Cognitive phenomenology* (pp. 174-196). Oxford, UK: Oxford University Press.
- Robertson, L. C. & Sagiv, N. (2005). *Synesthesia: Perspectives from cognitive neuroscience*. Oxford, UK: Oxford University Press.
- Schwitzgebel, E. (2002). How well do we know our own conscious experience? The case of visual imagery. *Journal of Consciousness Studies*, 9 (5-6), 35-53.
- (2008). The unreliability of naïve introspection. *Philosophical Review*, 117 (2), 245-273. [10.1215/00318108-2007-037](https://doi.org/10.1215/00318108-2007-037)
- (2011). *Perplexities of consciousness*. Cambridge, MA: MIT Press.

- (2013). Reply to Kriegel, Smithies, and Spener. *Philosophical Studies*, 165 (3), 1195-1206. [10.1007/s11098-013-0152-4](https://doi.org/10.1007/s11098-013-0152-4)
- Siegel, S. (2006). Which properties are represented in perception? In T. S. Gendler & J. Hawthorne (Eds.) *Perceptual experience* (pp. 481-503). Oxford, UK: Oxford University Press.
- Siewert, C. (2007). Who's afraid of phenomenological disputes? *Southern Journal of Philosophy*, 45 (S1), 1-21. [10.1111/j.2041-6962.2007.tb00107.x](https://doi.org/10.1111/j.2041-6962.2007.tb00107.x)
- (2011). Phenomenal thought. In T. Bayne & M. Montague (Eds.) *Cognitive phenomenology* (pp. 236-267). Oxford, UK: Oxford University Press.
- Smithies, D. (2012). A simple theory of introspection. In D. Smithies & D. Stoljar (Eds.) *Introspection and consciousness* (pp. 259-293). Oxford, UK: Oxford University Press.
- (2013). The nature of cognitive phenomenology. *Philosophy Compass*, 8 (8), 744-754. [10.1111/phc3.12053](https://doi.org/10.1111/phc3.12053)
- Spener, M. (2011a). Using first-person data about consciousness. *Journal of Consciousness Studies*, 18 (1), 165-179.
- (2011b). Disagreement about cognitive phenomenology. In T. Bayne & M. Montague (Eds.) *Cognitive phenomenology* (pp. 268-284). Oxford, UK: Oxford University Press.
- (2013). Moderate scepticism about introspection. *Philosophical Studies*, 165, 1187-1194.
- Spener, M. (Unpublished). *Calibrating introspection*.
- Tye, M. (1996). The function of consciousness. *Noûs*, 30 (3), 287-305. [10.2307/2216271](https://doi.org/10.2307/2216271)

“I just knew that!”: Intuitions as Scaffolded or Freestanding Judgements

A Commentary on Tim Bayne

Maximilian H. Engel

How reliable are intuitive or introspective judgments? This question has produced lively debates in two respective discussions. In this commentary I will try to show that the two phenomena of introspective and intuitive judgments are very closely related, so that the two separate philosophical debates about them can substantially inform each other. In particular, the intuition debate can profit from conceptual tools that have already been introduced to discussions about the reliability of introspection. Especially the distinction between scaffolded and freestanding judgements, which has been developed by [Tim Bayne & Maja Spener \(2010\)](#), can be used to more carefully investigate intuitions with respect to their epistemic reliability. After briefly applying this framework to some paradigm cases of “philosophically interesting” intuitions, I will come to the conclusion that most of these must be regarded as freestanding judgments and thus cannot play the role of reliable sources of evidence that they are supposed to play in some discussions in contemporary epistemology and methodology.

Keywords

Epistemic reliability | Experimental philosophy | Global pessimism | Local pessimism | Phenomenology of certainty | Philosophical intuitions | Scaffolded vs. freestanding judgments | Thought experiments

Commentator

[Maximilian H. Engel](#)
M.H.Engel.1@student.rug.nl
Rijksuniversiteit Groningen
Groningen, Netherlands

Target Author

[Tim Bayne](#)
tim.bayne@manchester.ac.uk
The University of Manchester
Manchester, United Kingdom

Editors

[Thomas Metzinger](#)
metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Introduction

What is the evidential status of introspective mental states? Can they be used as a source of knowledge like other classical candidates, e.g. experimental data, induction, or visual perception? Over the last few decades these questions have been addressed in philosophy of mind and epistemology in particular.¹ While on the one

hand optimists consider the wide-ranging use of introspection in philosophical debates unproblematic, pessimists on the other hand are very skeptical about the same subject matter. But how far can their skepticism go? Is it really the case that introspective insights are not only sometimes misleading, but generally false? These are the questions underpinning Tim Bayne’s article “Introspective Insecurity”. Here Bayne argues that a total dismissal of introspection as a tool for gaining information about

¹ In fact, [Eric Schwitzgebel \(2008\)](#) points out that there is a new trend of relying on introspection, even though this method itself is not new and its disadvantages were pointed out with the failure of introspective psychology at the beginning of the 20th century (c.f. [Lyons 1986](#)).

one's own conscious states (global pessimism) would not only be tremendously hard to imagine, but is also not warranted by the arguments raised in favour of that position. What these pessimistic arguments show, however, is that not all kinds of introspection can be used without thorough examination of their truth-tracking capacities. The resulting milder form of skepticism is what Bayne calls local pessimism. This distinction is what I consider Bayne's most important contribution to the introspection debate, because it helps to avoid an overhasty dismissal of a source of information that is used widely, not only in theorizing, but also in everyday life. He points out that what the global skeptic is missing is the idea that there are different kinds of introspective judgments, where not all are equally insecure. To distinguish between more secure cases of introspection and less secure ones, Bayne emphasizes a distinction introduced by him and his colleague Maja Spener in their paper [Introspective Humility \(2010\)](#), namely that of scaffolded versus freestanding judgments. While scaffolded judgments about one's introspective states are quite reliable, because their contents match closely with the contents of the non-introspective processes at work (e.g., visual experience), freestanding judgments lack this sort of reliability due to their abstract character. Simply put, the contents of freestanding judgments lack the close connection to what one wants to find out about the world or one's own mental states.

Another prominent, but also controversial candidate for being an epistemically useful source of evidence is intuition. Much like in the case of introspection, there is a large debate about the reliability and usefulness of intuitions in philosophical theorizing. This debate not only concerns epistemology and philosophy of mind, but also methodology, since many people claim that what philosophy does at its core is conceptual work on the basis of our rational (or conceptual) intuitions ([Bealer 1997](#); [Goldman 2007](#)). In the last few years, however, this idea of how to do philosophy has been harshly criticized from many different perspectives. While proponents of the fairly new project called experimental philosophy have tried to investigate the reliability of intuitions by con-

ducting survey studies collecting lay intuitions ([Weinberg et al. 2008](#); [Knobe 2007](#)), others have even gone so far as to argue that we do not use any intuitions at all in philosophical theorizing ([Cappelen 2012](#)). In any case, it is still open to debate whether intuitions can be used as reliable sources of evidence or not. Here I will first argue that this debate can be substantially informed by Bayne and Spener's idea of scaffolded versus freestanding judgments; this will be referred to as the Scaffolded vs. Freestanding Intuitions Thesis (SFIT). I will try to show that this is the case by highlighting some close connections and similarities between intuitions and introspection. Second, I will argue that in fact intuitions are often made accessible to the debates by introspection, namely in form of introspective insight about one's own private concepts.² This will be called the Introspection of Private Concepts View (IPCVC). Thereafter I will make my third claim, namely that many intuitions, at least those relevant in the debates in epistemology and methodology, are best regarded as freestanding judgments and thus should not count as reliable sources of evidence in philosophical debates. This third and last claim will be what I call the Unreliable Freestanding Intuitions Thesis (UFIT). As in the case of introspection, a total dismissal of intuitions is not (yet) warranted, but neither is their wide-ranging use in contemporary methodology. By applying Bayne's framework, i.e., the distinction between scaffolded vs. freestanding judgments, to the phenomenon of intuitive judgments, I will try to use this new conceptual tool to find a possible answer to the question of which kinds of intuitions are trustworthy and which should not be considered as reliable in philosophical debates.

2 Some connections and similarities between intuition and introspection

If one takes a look at the literature on introspection, one can find many metaphors that are derived from visual perception, i.e. that describe

² While in this commentary I will only concentrate on the influence of introspection on intuitive judgments, it is also worth noticing that both phenomena can also influence each other in the opposite direction. One factor that makes introspective insights feel so reliable at first glance is their intuitiveness. This would be a case in which intuition influences introspection.

the phenomenon as a sort of peering into one's own consciousness,³ as well as direct comparisons with visual perception, i.e., stating that the evidential status of introspection is or should be on a par with seeing the outside world. For example, in his depiction of the central idea behind optimism towards introspection, Bayne says that:

Roughly speaking, to regard introspection as able to *reveal* the phenomenal character of one's conscious states is to have an optimistic attitude towards it. (Bayne [this collection](#), my italics)

Or take Schwitzgebel, who, in his arguments against the accuracy of introspection, assesses the phenomenon by the standards of visual perception:

Does introspection reveal it to you *as clearly as visual observation* reveals the presence of the text before your eyes? Can you discern its gross and fine features through introspection as easily and confidently as you can, through vision, discern the gross and fine features of nearby external objects? (2008, my italics)

If one compares this to intuitions, one can see that they are treated in almost the same way. Here, the most prominent historical root of this equal treatment of not only intuitions and perception, but also intuitions and introspection, might be the work of John Locke, who at the beginning of the fourth book of his [Essay Concerning Human Understanding](#) states that all knowledge is at its core introspective and intuitive and can thus be regarded as the perception of agreement or disagreement between two ideas:

Knowledge then seems to me to be nothing but *the perception of connexion and agreement, or disagreement and repugnancy of any of our Ideas*. In this alone it consists.

³ A further hint at the equal treatment of introspection is the Latin origin of the term 'introspicere', which can be translated as 'to examine' or 'to look into'.

Where this Perception is, there is Knowledge, and where it is not, there, though we may fancy, guess, or believe, yet we always come short of Knowledge. (1975, p. 525, italics in the original)

But contemporary discussions concerning intuitions also suggest a similarity to perception. Take for example this short description by Ernest Sosa:

Intuition gives us *direct insight* into the general and abstract. (1998; my italics)⁴

For George Bealer, who is maybe the most radical proponent of an intuition-based philosophical methodology, the two phenomena are so closely related that he mentions them both as equal sources of evidence in philosophical theorizing:

So in this terminology, the standard justificatory procedure counts as evidence, not only *experiences, observations*, and testimony, but also intuitions. [...] When one has an intuition, however, often one is *introspectively aware* that one is having that intuition. On such an occasion, one would then have a bit of *introspective evidence* as well, namely, that one is having that intuition. (1997, my italics)

This similarity in the way of speaking about the two phenomena and their obvious entanglement in the debate about what counts as evidence⁵ gives us information about the explananda themselves. Both intuition and introspection can be consciously experienced by the subject that uses them to make a judgement.⁶ Furthermore, they are judged to be epistemically unproblematic, because the subject has direct access to them.

⁴ Here again the Latin origin 'intueri', which can be translated as 'to view' or also as 'to examine', underlines not only the folk psychological connection between intuition and perception but also the similarity between introspection and intuition.

⁵ For a general discussion of what counts, or should count as evidence, see [Williamson \(2007\)](#).

⁶ This does not mean that one always deliberately introspects or intuitively. This would be trivially false ([Sosa 1998](#)). What is meant is that one can in principle guide one's attention to the relevant mental state if necessary.

A good example is a classical Gettier-style intuition, such as “It *simply seems to me* that the person in that scenario does know that she is getting the job” (Gettier 1963). Not only the immediate reaction to Gettier cases, but also the way in which Gettier’s conclusion (i.e. that his thought experiments show that justified true belief does not sufficiently describe knowledge) were widely accepted among philosophers indicates that intuitive judgements are treated as unproblematic and reliable. The same holds for introspective judgements that do not only occur in philosophical debates but also in everyday-life belief formation. An example of such a belief could be expressed by a sentence like: “I surely can’t be mistaken in believing that I am consciously experiencing a red object in front of me at this very moment.” In the same way as in the case of intuitions, the results of introspection do not seem to require further questioning. In short, the act of introspecting something and the act of intuiting something both have a phenomenal aspect that makes them appear epistemically secure. In the course of this commentary this aspect will be referred to as a phenomenology of certainty.⁷ In fact, I would say that this phenomenal aspect is the reason why the introspection as well as the intuition debate are as controversial as they are. Both phenomena come at first glance with a seeming of epistemic security (or even infallibility), and only after close examination are some insecurities revealed. This phenomenology of certainty, however, does not immediately show that intuitions and introspection inform a subject securely about the truth of a matter. My introspective judgment about the what-it-is-likeness of understanding a sentence in a foreign language or my intuitive judgment about whether a person has knowledge or not are always in need of further justification. It would be a very hasty step to go from the phenomenology of certainty to full-fledged certainty (Metzinger & Windt 2014).

⁷ It is important to notice that the “phenomenology of certainty” presupposes a “phenomenology of knowing”. This is best regarded as the “phenomenology of knowing that one knows”. For my purposes here the “phenomenology of knowing”, though important, is not the interesting phenomenal aspect of intuitions or introspective insights. I hold the “phenomenology of certainty” far more interesting, because I think that it is that phenomenology that leads to the strong sense of infallibility of intuitive, as well as introspective judgments.

So then what can the two phenomena inform a subject about? The least controversial description of what introspective states are would be along the lines of (Schwitzgebel’s description:

A word about ‘introspection’. I happen to regard it as a species of attention to currently ongoing conscious experience, but I won’t defend that view here. The project at hand stands or falls quite independently. Think of introspection as you will—as long as it is the primary method by which we normally reach judgments about our experience in cases of the sort I’ll describe.⁸ (2008)

Thus construed, introspection mainly informs a subject about the qualitative aspects of her experience. Simply put, what we do when we introspect is to pay attention to the what-it-is-likeness of our experience.⁹ This aspect of experience, however, is extremely subjective and private. It is (if even possible) not easy to arrive at scientifically informative generalizations¹⁰ from such subjective data.¹¹ What is needed to secure information of that kind is the right kind of embeddedness in other, more secure ways of gaining knowledge about a subject matter. Such judgments about a subject’s experience are what Bayne and Spener, at least by the way I understand them, refer to as scaffolded judgments (2010; Bayne this collection). For ex-

⁸ The cases he describes in that paper are from the same domains of experience that Bayne discusses in his article for this volume namely emotion, visual perception, and cognitive phenomenology.

⁹ Note that due to restrictions of space I will cover only, the most relevant interpretation of introspection, which can be described as a sort of inward perception. The word “perception” here is to be read in a metaphorical way. It is not meant to express a commitment to something along the lines of a higher-order perception view on introspection (Güzeldere 1995). Rather this inward “perception” can be understood as kind metacognition that helps a subject to conceptualize her own experiences. For a more detailed distinction between different kinds and qualities of introspection, see Metzinger (2003, p. 35).

¹⁰ Though this might not be a problem for relying on introspection in the case of perception, it becomes more pressing when it comes to using introspective data to inform epistemology or methodology.

¹¹ A further methodological problem that needs to be taken into consideration is the fact that when collecting data about introspective or intuitive states one has to rely on a subject’s report about the relevant mental state. This can be a possible source of contamination, which makes an investigation of the phenomena even more difficult (Cummins 1998).

ample, my introspective judgment about my red experience is not exhaustively justified by itself, but by the close match of the content of my introspective state and the non-phenomenal aspects of my visual observation. Only then can introspection play an evidential role,¹² and thus contribute to knowledge about one's own conscious states. But what if there is no such match? If introspection is concerned with more abstract contents, like, for example, the basic structures of intentionality or thought in general, the lack of embeddedness at least increases the insecurity of the judgment and thus makes it an unreliable source of knowledge. Judgments of that kind, again following Bayne and Spencer, are called freestanding judgments.

Let us now turn to intuitions. What are intuitions about? First of all, it is important to say that not all kinds of intuitions are relevant to philosophical debates. Cases of intuitive controls on a smartphone, for example, are not at the core of the debate. What is meant by philosophically interesting intuitions can be most appropriately expressed by the term conceptual intuition. In short, intuitions in a philosophically relevant sense are judgments that are shaped by the concepts a person has of some subject matter or phenomenon. Usually those intuitions are tested by conducting thought experiments in which a case is described that should (or should not) fulfil all necessary and sufficient conditions of a concept. Then one is supposed to take that very concept and check if it applies to the case (or not). This is why Alvin Goldman also refers to philosophical intuitions as “application intuitions” (2007). Probably the most prominent examples of such intuition-testing thought experiments are Gettier cases. Going back to Edmund Gettier's famous paper, Gettier cases describe scenarios in which a person appears to lack knowledge, despite the fact that the classical conditions for having knowledge, namely, having a justified true belief, are met (1963). But can these conceptual intuitions in fact inform us about what knowledge is in general, or do those cases simply inform us about our personal concepts? Findings from the

fairly new field of experimental philosophy, though highly controversial (Cullen 2010), indicate that conceptual intuitions that have been treated as general intuitions, like those in Gettier cases, are in fact highly idiosyncratic, and thus it is still an open question whether they can lead to generalizations about the concept at hand (Alexander 2012).¹³ In other words, one could argue that conceptual intuitions are the reflections of a subject's *idiosyncratic history of concept acquisition* (Bieri 2007).

So intuitions—or more precisely their contents—reflect upon a person's individual, highly subjective concepts. Just like in the case of introspection (which has been shown to be very subjective as well), we need to investigate whether it is possible to move from those personal concepts to general claims about their contents in a reliable way.

I take all of the above-mentioned similarities between introspection and intuition to be sufficient for investigating the reliability of intuitions with conceptual tools and insights that have already been introduced and established to the introspection debate. Thus, I will in the next section try to clarify what counts as an epistemically reliable intuition by applying the distinction between scaffolded and freestanding judgments from the introspection debate to intuitions. In other words, I will investigate intuitions as scaffolded vs. freestanding intuitions (SFIT).

SFIT =_{Df} Due to the similarities between introspection and intuition, one can also distinguish between scaffolded and freestanding intuitions.

3 Philosophical intuitions as freestanding judgments

Before we examine whether philosophical intuitions are best understood as scaffolded or freestanding judgments, it will be helpful to

¹³ In addition to these findings, it is also an advantage of treating intuitions as reflections on personal concepts, because such a view is likely to be naturalized (Goldman 2007). Arguments from obscurity or empirical implausibility of the type that have been raised against other construals of intuition, such as Platonic insights into the laws of nature (Brendel 2004), can thus be avoided.

¹² Even if this role is then obviously a minor one in forming a belief about the world.

take a closer look at how intuitions are treated in philosophical theorizing. For this we go back again to the paradigm case of intuition-based philosophy: Gettier (1963) cases. The expected (and therefore long unchallenged) outcome of those thought experiments is that the person reflecting on the cases admits that they describe instances of justified true belief that at the same time fail to count as knowledge. How do we know they're not knowledge? We just know! Reflecting on that answer one can come to the conclusion that one has an intuition about the concept of knowledge. The next question that then needs to be answered is how a person arrives at that conclusion. I claim that this is done by introspection. As described above, introspection is best understood as the act of paying attention to one's conscious states of experience, or in other words about the phenomenal aspects of experience. In the case of an intuition, this phenomenal aspect would be the above-mentioned phenomenology of certainty. To summarize this, conceptual intuitions are reflected upon by introspecting on one's own concepts and their applicability conditions. This practice is what I call the Introspection of Personal Concepts View of Intuitions (IPCV).

IPCV =_{Df} Conceptual Intuitions are made accessible by introspecting one's own phenomenology of certainty towards the applicability of a certain concept.

Following IPCV, this practice is then of course vulnerable to the same skeptical challenges that have been raised against introspection in general. How accurately can I introspect what constitutes my concept of knowledge? What about modal aspects like the necessity of a proposition? These questions can be made more accessible by thinking about intuitions in terms of scaffolded or freestanding judgments.

Again taking the Gettier intuition about knowledge, what makes this intuition, even though not universal, so astonishingly stable among Western philosophers? I argue that this is due to the close match between the content of the intuition (i.e. "She doesn't know!") and the rules that one learns about how to successfully use the concept of

knowledge in our cultural niche (i.e.: "Only ascribe knowledge if a person is justified in the right way to believe a proposition!").¹⁴ So in the context of Western philosophy, the intuitive judgment can be regarded as a scaffolded and thus reliable judgment. It is reliable because it is embedded in our conventional, everyday use of the word "knowledge".¹⁵ But what about knowledge in general, i.e., outside the context of Western culture? In that case, the content of the intuition, due to its personal character, would not match the context-free, abstract use of the concept of knowledge. The judgment would be a freestanding judgment and thus an unreliable source of evidence for making general claims about knowledge. This would perfectly fit the idea of intuitions as individually-acquired concepts and also explain findings from experimental philosophy, which indicate that intuitions are highly variable among different cultures (Weinberg et al. 2008). One could now argue that, even if I am correct about conceptual intuitions like those in Gettier cases, there are basic intuitions that are reliable. A candidate for such an intuition is presented by Bealer in the form of rational intuitions:

By contrast, when we have a rational intuition—say, if P then not not P —it presents itself as necessary: it does not seem to us that things could be otherwise; it must be that if P then not not P . (I am unsure how exactly to analyze what is meant by saying that a rational intuition presents itself as necessary. Perhaps something like this: necessarily, if x intuits that P , it seems to x that P and also that necessarily P [...].) (1997)

The reliability of such a basic intuition can also be accommodated in the terminology of scaffolded and freestanding judgments. Due to the close match between our intuition and the way

¹⁴ Surely this is a very simplified and rough description of concept acquisition. Further details should be empirically investigated, but due to limited space, and for and the purposes of my argument, this must suffice.

¹⁵ The scaffold here would be the proper use of a word or concept in its respective culture or context. Further, notice that it is also possible to have several types of scaffolding at the same time, like conceptual expertise (i.e. cases in which a person has a significant amount of background knowledge about special concept) plus the above-mentioned cultural scaffolding. For a defence of conceptual expertise, see Williamson (2011).

in which we learned to describe the world, in which it never is the case that p while simultaneously not p , we can regard that intuition as a scaffolded judgment. Concerning the intuition about the necessity of this intuited content, however, the personal character of intuitions again does not warrant the generalization. Statements about the modal status of the claim are perhaps secured by correctly applying the laws of logic (like in the above mention example of the principle of contradiction), but not by my personal intuition (Alexander 2012; Pust 2014). But even if this is true and thus if such basic intuitions are always reliable, it still needs to be shown by general optimists, concerning the reliability of intuitions, how this extends to more complex phenomena like those often discussed in the intuition debate (Cappelen 2012). I take the above-discussed cases of Gettier-intuitions and Bealer's rational intuitions as evidence that we should at least doubt that most intuitions that are taken as reliable sources of evidence are sufficiently scaffolded. Until this is shown I would advise that we stay skeptical and regard those intuitions as Unreliable Freestanding Intuitions (UFIT).

UFIT =_{DF} Many intuitions that are treated as reliable sources of evidence in philosophical theorizing lack the right scaffolding and must thus be regarded as freestanding intuitions, which makes them epistemically unreliable.

4 Conclusion

In this commentary I have tried to show that the connections between introspection and intuitions are so profound that the debates about the two phenomena can inform each other substantially, and in particular how ideas from the introspection debate can help to clarify open questions in the intuition debate (SFIT). I have taken the idea of scaffolded and freestanding judgments from the introspection debate and applied it to that about intuitions. In so doing, I have tried to show that the wide-ranging skepticism about introspection also concerns intuitions, since many intuitions are investigated by

introspecting on one's phenomenology of certainty that typically accompanies intuitions, as well as introspection itself (IPCV). Bayne's introduction of the scaffolded versus freestanding judgments idea suggests that a global pessimism towards introspection is not warranted by the arguments that are raised by proponents of such a position. I hope to have shown that the same is true in the case of intuitions, which can also be reliable if they are embedded in the right context, or if concerning the basic structures of our experience. The question for further discussion has now become how big the scope of both scaffolded introspective and scaffolded intuitive judgments actually is. Is it possible to develop clear-cut criteria for when a content is sufficiently scaffolded? Must one draw further distinctions and introduce different kinds, or at least a gradual concept, of scaffolding? So far, applied to often very abstract epistemic targets in philosophy, my predictions for the scope of scaffolded judgments in the on-going debates are not very optimistic. I would advise that without further argumentation for the scaffolding of abstract intuitions they are best regarded as freestanding judgments (UFIT). I agree with Sosa when he says, about the skeptical challenges to intuitions: "If that sort of consideration is a serious indictment of intuition, therefore, it seems no less serious when applied to introspection [...]" (1998). The only difference might be that I hold this to be bad news for proponents of the widespread use of both phenomena, rather than a convincing defence of their general reliability.

Acknowledgements

I would like to thank Tim Bayne for his inspiring target article, as well as two anonymous reviewers for their very helpful suggestions and critical remarks. I would also like to thank Thomas Metzinger and Jennifer Windt for the opportunity to contribute to this collection.

References

- Alexander, J. (2012). *Experimental philosophy: An introduction*. Cambridge, UK: Polity Press.
- Bayne, T. (2015). Introspective Insecurity. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M.: Mind Group.
- Bayne, T. & Spener, M. (2010). Introspective humility. *Philosophical Issues*, 20 (1), 1-22.
[10.1111/j.1533-6077.2010.00176.x](https://doi.org/10.1111/j.1533-6077.2010.00176.x)
- Bealer, G. (1997). Intuition and the autonomy of philosophy. In M. de Paul & W. Ramsey (Eds.) *Rethinking intuition: The psychology of intuition and its role in philosophical inquiry* (pp. 201-239). Lanham, MD: Rowman & Littlefield Publishers.
- Bieri, P. (2007). Was bleibt von der analytischen Philosophie? *Deutsche Zeitschrift für Philosophie*, 55 (3), 333-344. [10.1524/dzph.2007.55.3.333](https://doi.org/10.1524/dzph.2007.55.3.333)
- Brendel, E. (2004). Intuition pumps and the proper use of thought experiments. *Dialectica*, 58 (1), 89-108.
[10.1111/j.1746-8361.2004.tb00293.x](https://doi.org/10.1111/j.1746-8361.2004.tb00293.x)
- Cappelen, H. (2012). *Philosophy without intuitions*. Oxford, UK: Oxford University Press.
- Cullen, S. (2010). Survey driven romanticism. *Review of Philosophy and Psychology*, 1 (2), 275-296.
[10.1007/s13164-009-0016-1](https://doi.org/10.1007/s13164-009-0016-1)
- Cummins, R. (1998). Reflection on reflective equilibrium. In M. DePaul & W. Ramsey (Eds.) *Rethinking intuition: The psychology of intuition and its role in philosophical inquiry* (pp. 113-127). Lanham, MD: Rowman & Littlefield Publishers.
- Gettier, E. (1963). Is justified true belief knowledge? *Analysis*, 23 (6), 121-123. [10.1093/analys/23.6.121](https://doi.org/10.1093/analys/23.6.121)
- Goldman, A. (2007). Philosophical intuitions: Their target, their source, and their epistemic status. *Grazer Philosophische Studien*, 74 (1), 1-26.
- Güzeldere, G. (1995). Is consciousness the perception of what passes in one's own mind? In T. Metzinger (Ed.) *Conscious experience* (pp. 335-357). Paderborn, GER: Schöningh.
- Knobe, J. (2007). Experimental philosophy. *Philosophy Compass*, 2 (1), 81-92.
[10.1111/j.1747-9991.2006.00050.x](https://doi.org/10.1111/j.1747-9991.2006.00050.x)
- Locke, J. (1975). *An essay concerning human understanding*. Oxford, UK: Oxford University Press.
- Lyons, W. (1986). *The disappearance of introspection*. Cambridge, MA: MIT Press.
- Metzinger, T. (2003). *Being no one: The self-model theory of subjectivity*. Cambridge, MA: MIT Press.
- Metzinger, T. & Windt, J. M. (2014). Die phänomenale Signatur des Wissens: Experimentelle Philosophie des Geistes mit oder ohne Intuitionen? In T. Grundmann, J. Horvath & J. Kipper (Eds.) *Die experimentelle Philosophie in der Diskussion* (pp. 279-321). Berlin, GER: Suhrkamp.
- Pust, J. (2014). Intuition. *The Stanford Encyclopedia of Philosophy*, spring E. N. Zalta (Ed.) <http://plato.stanford.edu/archives/spr2014/entries/intuition/>
- Schwitzgebel, E. (2008). The unreliability of naïve introspection. *Philosophical Review*, 117 (2), 245-273.
[10.1215/00318108-2007-037](https://doi.org/10.1215/00318108-2007-037)
- Sosa, E. (1998). Minimal intuition. In M. de Paul & W. Ramsey (Eds.) *Rethinking intuition: The psychology of intuition and its role in philosophical inquiry* (pp. 257-269). Lanham, MD: Rowman & Littlefield Publishers.
- Weinberg, J., Nichols, S. & Stich, S. (2008). Normativity and epistemic intuitions. In J. Knobe & S. Nichols (Eds.) *Experimental philosophy* (pp. 17-45). Oxford, UK: Oxford University Press.
- Williamson, T. (2007). *The philosophy of philosophy*. Malden, MA: Blackwell Publishing.
- (2011). Philosophical expertise and the burden of proof. *Metaphilosophy*, 42 (3), 215-229.
[10.1111/j.1467-9973.2011.01685.x](https://doi.org/10.1111/j.1467-9973.2011.01685.x)

Introspection and Intuition

A Reply to Maximilian H. Engel

Tim Bayne

This paper is a response to Maximilian H. Engel's commentary on my target paper, in which I provided a critical examination of pessimism accounts of the trustworthiness of introspection. Engel's focuses on the distinction that I drew between two kinds of introspective judgments, scaffolded judgments and freestanding judgments, and suggests that this distinction might fruitfully illuminate the epistemology of intuitive judgments. I present some doubts about whether the distinction can be transferred to intuition in this way, and also sketch a more fundamental contrast between introspective judgments and intuitive judgments.

Keywords

Free-standing judgments | Introspection | Intuition | Scaffolded judgments

Author

[Tim Bayne](#)

tim.bayne@manchester.ac.uk
The University of Manchester
Manchester, United Kingdom

Commentator

[Maximilian H. Engel](#)

M.H.Engel.1@student.rug.nl
Rijksuniversiteit Groningen
Groningen, Netherlands

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Introduction

Let me begin by thanking Maximilian H. Engel for his commentary. I take the heart of his paper to consist in the suggestion that the distinction between freestanding and scaffolded judgments which Maja Spener and I ([Bayne & Spener 2010](#)) developed in connection with introspection can be usefully applied to the epistemology of intuition. I will start by revisiting the freestanding/scaffolded distinction, before turning to Engel's proposal.

The epistemology of introspection is that it is not flat but contains peaks of epistemic security alongside troughs of epistemic insecurity. Any attempt to understand the epistemology of

introspection needs to take this landscape into account, for although our pretheoretical views concerning the epistemology of introspection are not sacrosanct they do form a useful constraint on theorizing about introspection. Any account of introspection should explain why some introspective judgments strike us as highly secure whereas others seem to be insecure.

This is where the distinction between scaffolded and freestanding judgments comes in. Both types of judgments have as their intentional objects current conscious states that one takes oneself to be in. (The notion could also be applied to judgements concerning the states

that one is not in.) An introspective judgment is scaffolded when the subject is disposed to make a first-order judgment whose content bears a rough correspondence to that of the introspective judgment. For example, the judgment that one is experiencing a red light in front of one is scaffolded by the disposition to judge that there is a red light in front of one, whereas there is no such first-order disposition corresponding to the introspective judgment that one is merely imagining or thinking about a red light. Experiences that are the intentional objects of scaffolded judgments are themselves employed in world-directed first-order judgments, whereas that is not the case where free-standing judgments are concerned. Contrary to what Engel suggests, there is no commitment here to the idea that only scaffolded judgments are epistemically trustworthy. The idea, rather, is that scaffolded judgments have a certain kind of first-person warrant that free-standing judgments tend to lack.

2 From introspection to intuition?

Engel argues that the distinction between scaffolded and free-standing judgments can also be applied to the kinds of judgments deployed in debates about philosophical intuitions, and also suggests that most such judgments—or at least, those which are of central philosophical interest—are best regarded as free-standing, and thus lack the kind of warrant that we might want for them.

Although I welcome Engel's attempt to extend the distinction between scaffolded and free-standing judgments beyond the domain of introspection, I am not convinced that it does much to illuminate the epistemology of intuition. The first issue that needs to be addressed is the fact that intuitive judgments don't form a single, well-behaved class. One kind of intuitive judgment that is of philosophical interest concerns the modal structure of the world, as when one judges that it is necessarily true that $2+2=4$ or that it is only contingently true that Aristotle was a philosopher. But as far as I can tell, Engel is not concerned with intuitive judgments of this kind, but with what we might call intuitions of concept ap-

plication. Such judgments are concerned with the question of whether a certain concept (such as <knowledge>) ought to be applied to a certain state of affairs.

In explaining how the contrast between scaffolded and free-standing judgments might apply to intuitive judgments Engel writes:

Again taking the intuition about knowledge, what makes this intuition, even though not universal, so astonishingly stable among Western philosophers? I argue that this is due to the close match between the content of the intuition (i.e. “she doesn't know!”) and the rules one learns to use [regarding] the concept of knowledge in our cultural niche (i.e.: “Only ascribe knowledge if a person is appropriately justified in believing a proposition!”). So in the context of Western philosophy, the intuitive judgment can be regarded as a scaffolded and thus reliable judgment. ([this collection](#), p. 6)

It is certainly true that an individual's use of a concept is scaffolded by the practices of the culture in which they are embedded. As Kant pointed out, we learn how to apply concepts by noting how they are applied by those around us. Kant (A134/B174) described examples as the “*Gängelwagen* of thought”, where a *Gängelwagen* is a walking frame or go-kart that is harnessed to an infant in order to help it learn to walk. But although this form of support is indeed a kind of scaffolding, it differs in important ways from the kind of scaffolding that I had in mind. In the sense of the term that Spener and I had in mind, a scaffolded judgment is a judgment that is underpinned by a disposition to make a first-order judgment whose content roughly corresponds to the content of the scaffolded judgments. As far as I can see, intuitive judgments are not scaffolded in this sense, in part because intuitive judgments are already “first-order”. So, although I would certainly agree that the possession of such concepts as <knowledge> is supported by one's cultural niche, it doesn't follow that the intuitive judgments about when it is and isn't appropriate to apply this concept are scaffolded.

3 Intuitive disagreement

In closing, let me mention an important background issue concerning which Engel and I appear to have different views. Engel, I take it, holds that the disagreement in intuitive judgments regarding concept application should be regarded as epistemically troublesome in much the way that disagreement about introspective judgment is regarded as epistemically troublesome. The idea is that in both cases there are objective facts of the matter, and the existence of widespread disagreement indicates that significant numbers of individuals are systematically mistaken about what those facts are.

Although I am inclined to accept this diagnosis when it comes to many introspective disagreements, I do not find it particularly plausible when it comes to disagreements concerning intuitions of concept application. Here's why. Suppose that Weinberg and his collaborators are right when they suggest that low-socioeconomic status individuals are disposed to apply the concept <knowledge> in contexts where high-socioeconomic status (SES) individuals are disposed to withhold it (Weinberg et al. 2001). Would it follow (as Engel seems to assume) that at least one of these groups is mistaken about a matter of objective fact? I don't think so. It seems to me more plausible to assume that low-SES subjects and high-SES subjects simply have different concepts (or "conceptions", if you prefer) of knowledge, and each of them is applying its own concept correctly. The two concepts are similar enough to be both associated with the single word "knowledge", but there is no case for regarding one of these concepts as superior to the other, or for thinking that only one of them truly captures the essence of knowledge. They are simply different concepts.

If this is right, then apparent disagreement between the judgments of low-SES subjects and high-SES subjects about whether or not S knows that P is not substantive in the way in which most introspective disagreement appears to be. Moreover, it seems to me that something similar should be said concerning many (if not all) disputes about the application of other central philosophical concepts. (One needs to

take the possibility of performance errors into account here, but such problems will typically be minimized in philosophical contexts.) But I wouldn't want to commit myself to this account of *all* intuitive disputes. In particular, it seems to me that introspective disputes concerning modal matters are likely to be substantive in a way in which disagreements about intuitions regarding concept application are not.

4 Conclusion

In his commentary Engel suggests that the contrast between scaffolded and freestanding judgments that Spener and I applied to introspection might also be usefully applied to intuition. Although I welcome Engel's attempt to extend the distinction between scaffolded and freestanding judgments beyond its original sphere of application, I have suggested that such a move might not be quite as straightforward as Engel takes it to be, for there don't appear to be any first-order judgments that might scaffold intuitive judgments in the way that first-order perceptual judgments scaffold certain kinds of introspective judgments. But although I cannot see how the distinction between scaffolded and freestanding judgments might apply to intuition, I certainly share Engel's conviction that "comparing and contrasting" the epistemology of introspection with that of intuition is a fruitful exercise, for both domains pose the puzzle of how we might reconcile individual certainty and apparent self-evidence with intersubjective disagreement.

References

- Bayne, T. & Spener, M. (2010). Introspective humility. *Philosophical Issues*, 20 (1), 1-22. [10.1111/j.1533-6077.2010.00176.x](https://doi.org/10.1111/j.1533-6077.2010.00176.x)
- Engel, M. H. (2015). "I just knew that!": Intuitions as scaffolded or freestanding judgements: A commentary on Tim Bayne. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Kant, I. (1781/1787). *The critique of pure reason*. Salt Lake City, UT: Project Gutenberg eBook.
- Weinberg, J., Nichols, S. & Stich, S. (2001). Normativity and epistemic intuitions. *Philosophical Topics*, 29 (1&2), 429-460. [10.5840/philtopics2001291/217](https://doi.org/10.5840/philtopics2001291/217)

Meaning, Context, and Background

Christian Beyer

It is widely held that (truth-conditional) meaning is context-dependent. According to John Searle's radical version of contextualism, the very notion of meaning "is only applicable relative to a set of [...] background assumptions" (Searle 1978, p. 207), or background know-how. In earlier work, I have developed a (moderately externalist) "neo-Husserlian" account of the context-dependence of meaning and intentional content, based on Husserl's semantics of indexicals. Starting from this semantics, which strongly resembles today's mainstream semantics (section 2) I describe the (radical) contextualist challenge that mainstream semantics and pragmatics face in view of the (re-)discovery of what Searle calls the background of meaning (section 3). Following this, and drawing upon both my own neo-Husserlian account and ideas from Emma Borg, Gareth Evans and Timothy Williamson, I sketch a strategy for meeting this challenge (section 4) and draw a social-epistemological picture that allows us to characterize meaning and content in a way that takes account of contextualist insights yet makes it necessary to tone down Searle's "hypothesis of the Background" (section 5).

Keywords

Background hypothesis | Borg | Content | Context | Contextualism | Evans | Externalism | Husserl | Intentionality | Interpretation | Knowledge | Meaning | Minimalism | Reference | Searle | Williamson

Author

Christian Beyer
christian.beyer@phil.uni-goettingen.de

Georg-August-Universität
Göttingen, Germany

Commentator

Anita Pacholik-Żuromska
anitapacholik@gmail.com

Uniwersytet Mikołaja Kopernika
Toruń, Poland

Editors

Thomas Metzinger
metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Introduction

"Meaning" is a popular term in philosophical slogans. Meaning is said to be normative; not to be in the head. The notion of meaning is (nevertheless) said to be the key to the notion of intentional content, to only be applicable relative to a set of background assumptions, and meaning is said to be context-dependent. These slogans are not unrelated, and all of them have a reading, I suppose, in which they are true. Here I shall mainly focus on the last two slogans, regarding background and context. My main question will be twofold:

1. In which sense, and to which extent, can the meaning of assertive utterances be said to be context-dependent?

2. Does this context-dependence have an impact on the validity of Searle's Background Hypothesis, which states that the intentional experiences expressed by assertive utterances, and bearing their respective meaning, and the mental acts of grasping this meaning, both require a non-intentional background on the part of the speaker/hearer, relative to which the truth-conditional content and the satisfaction conditions of the relevant experience are determined?

The upshot will be that (1) whilst there may be expressions lacking the context-sensitivity that many expressions (namely, the indexicals) possess in virtue of their conventional linguistic

meaning, there is a sense (to be explained in terms of the background) in which context-dependence is ubiquitous; but that (2) this context-dependence does not prevent competent language users who lack the sort of individual background in terms of which this particular context-dependence can be defined (the “consumers”) from grasping the literal truth-conditional meaning (the semantic content) which an assertive utterance expresses on a given occasion.

2 Three levels of meaning

An early proponent of the view that meaning is context-dependent is Husserl. His thought on meaning, as manifested in his first major work [Logical Investigations](#), starts out from the problem of what it is for a linguistic expression, as used by a speaker or (scientific) author, to function as a meaningful unit.¹

Husserl’s approach is to study the units of *consciousness* that the respective speaker deliberately presents herself as having—that she “intimates” or “gives voice to”—when expressing the meaning in question. This is what Searle refers to as the condition of sincerity of the relevant speech act ([Searle 1983](#), pp. 9-10). These units of consciousness Husserl labels INTENTIONAL EXPERIENCES or ACTS, since they always represent something—thus exhibiting what Brentano called intentionality. They are “about”, or “as of,” something. For instance, if you claim “One of my goals is to defend contextualism,” you give voice to a judgment or belief-state to the effect that defending contextualism is among your goals. This judgment is intentional, in that it represents a state of affairs, namely your having a particular goal; it is “about” that state of affairs, even if the latter does not exist (i.e., obtain) because you do not have that goal. Now it is the content of this judgment (which may be empty or unfulfilled, i.e., made in the absence of a corresponding intuition, such as a corresponding perception) that a hearer has to know in order to understand your utterance, i.e., to grasp its literal meaning. Thus, the (unfulfilled) judgment functions as the “meaning-

bestowing” or “meaning conferring act” ([Husserl 2001](#), p. 192) regarding the sentence uttered. This act is given voice to, or intimated, “in the narrow sense” ([Husserl 2001](#), p. 189)—it is the condition of sincerity of the speech act. However, in the present example (“One of my goals is to defend contextualism”) the speaker also deliberately presents herself as someone who wants to defend contextualism; after all, she explicitly ascribes that intention to herself. This latter act (the intention in question) is given voice to “in the broader sense” only ([Husserl 2001](#), p. 189), as it fails to be the meaning-bestowing act regarding the sentence uttered and thus to be given voice to in the narrow sense. In other words, the speaker intentionally presents herself as performing or undergoing that act, but if the hearer does not recognize that intention he does not *thereby* fail to grasp the literal truth-conditional meaning of the utterance. Again, if you assert “This is a blooming tree,” you give voice, in the narrow sense, to a demonstrative judgment; but you also present yourself as perceiving (or having perceived) something as a blooming tree, where the act of perception is given voice to in the broader sense. This perceptual act verifies the unfulfilled judgment by intuitively “fulfilling” it ([Husserl 2001](#), p. 192). Since the meaning-bestowing act finds its aim, so to speak, in this intuitive fulfilment, Husserl also refers to it as the corresponding “meaning intention” ([Husserl 2001](#), p. 192). Since any meaning intention aims at its intuitive fulfilment, every meaningful utterance can in principle be made to give voice (in the broader sense) to such an act of fulfilment, provided its literal meaning is not evidently inconsistent. In sections 3 and 4 I shall argue that only the group of speakers capable of both making and understanding such epistemic implicatures (the “producers”) must meet the requirements of Searle’s [Background Hypothesis](#). One does not have to meet these requirements in order to express, or correctly ascribe, a meaning intention and thus grasp the literal truth-conditional meaning of an (assertive) utterance.

The “original function” of linguistic expressions is their communicative use in giving voice to meaning-bestowing acts, or meaning intentions ([Husserl 2001](#), p. 189). However, this “indicating (*anzeigende*)” function is not essen-

¹ For the following presentation of Husserl’s theory of meaning cf. Beyer & Weichold 2011, p. 406.

tial to their functioning as meaningful units, as they can also be employed “in [the] solitary life [of the soul] (*im einsamen Seelenleben*),” thanks to meaning-bestowing acts not actually given voice to but experienced all the same (Husserl 2001, pp. 190-191). But these acts and the meanings they bear are constrained by semantic factors concerning the linguistic expressions employed, with these factors being determined by linguistic conventions regarding the relationship between their meaning and the features of non-linguistic reality they serve to represent:

[...] it pertains to the *usual* [i.e., conventional; CB] sense of these classes of expressions, that they owe their determinate meaning to the occasion [...] [T]heir [respective] meaning is oriented in each case to the individual instance, though the manner of this orientation is a matter of usage.² (Husserl 2001, p. 221)

Husserl’s theory of meaning strongly resembles the mainstream view in philosophy of language attacked by Searle and other contextualists. In the following passage Searle gives a concise summary of that view:

Sentences have literal meanings. The literal meaning of a sentence is entirely determined by the meanings of its component words (or morphemes) and the syntactical rules according to which these elements are combined. [...] The literal meaning of a sentence needs to be sharply distinguished from what a speaker means by the sentence when he utters it to perform a speech act [...]. For example, in uttering a sentence a speaker may mean something different from what the sentence means, as in the case of metaphor; or he may even mean the opposite of what the sentence

means, as in the case of irony; or he may mean what the sentence means but mean something more as well, as in the case of conversational implications and indirect speech acts. [...] For sentences in the indicative, the meaning of the sentence determines a set of truth conditions [...] Sometimes the meaning of a sentence is such that its truth conditions will vary systematically with the contexts of its literal utterance. Thus the sentence ‘I am hungry’ might be uttered by one person on one occasion to make a true statement and yet be uttered by another person, or by the same person on another occasion, to make a false statement. [...] It is important to notice however that the notion of the meaning of a sentence is absolutely context free. Even in the case of indexical sentences the meaning does not change from context to context; rather the constant meaning is such that it determines a set of truth conditions only relative to a context of utterance.³ (Searle 1978, pp. 207-208)

To bring out the relevant semantic factors, consider what Husserl calls “essentially occasional expressions,” i.e., systematically context-sensitive, or indexical, expressions such as “I,” “here,” “now,” “I am here now.”⁴ In his pioneering discussion of these expressions in the first *Logical Investigation*, paragraph 26, Husserl introduces the semantic distinction between, on the one hand, an expression’s *general meaning-function* (i.e., the linguistic meaning of the expression, roughly corresponding to what Kaplan calls “character”) and, on the other hand, the propositional, or sub-propositional,⁵ content – the “*respective meaning*” – expressed in a given context of utterance (Husserl 2001, p. 218). If, for example, you and I both say “I,” then our two

2 The German original runs: “Es gehört zur *usuellen* Bedeutung dieser Klassen von Ausdrücken, ihre Bedeutungsbestimmtheit erst der Gelegenheit zu verdanken [...] [Sie orientieren] ihre jeweilige Bedeutung erst nach dem Einzelfall, während doch die Weise, in der sie dies tun, eine usuell ist.” (Hua XIX/1, pp. 91f.) So Husserl does not subscribe to a Humpty-Dumpty view of meaning, according to which the meaning of an expression in the mouth of a speaker is solely determined by what the speaker wants the expression to mean on the respective occasion; cf. Beyer 2000, pp. 78-79.

3 For an overview of more recent developments in semantics and pragmatics, cf. Lepore & Smith 2006, and the entries in Barber & Stainton 2010.

4 Unlike mainstream semantics, Husserl considers such expressions to be ubiquitous in empirical thought and speech; cf. Husserl 2001, p. 7. The approach to meaning I shall sketch below supports this contention.

5 A sub-propositional content is a non-propositional content (or respective meaning) that is a subpart of a propositional content. Singular and general terms may be used to express sub-propositional contents.

utterances share the same general-meaning function but express different respective meanings, with different referents. Again, if you and I both assert “I have blood type A,” our utterances share the same general meaning-function but express different respective meanings, with different truth conditions. These respective meanings, or truth-conditional contents, are often referred to as *propositions* expressed by the utterance of a sentence.

Husserl regards the general meaning-function as fixed by common usage (Husserl 2001, p. 221). The *respective* meaning determines the expression’s reference, or truth condition, in the sense that two expressions sharing that meaning are thus bound to refer to the same object(s), or to represent the same state of affairs, if any. Husserl construes “respective meanings” as two-factored, with the general meaning function plus the relevant context of utterance (if any) determining the meaning in question. Thus we have two levels of meaning⁶ being expressed when a meaning intention is given voice to:

General meaning-function (conventional linguistic meaning, “character”) =_{DF} The general meaning-function of an expression is a function yielding a respective meaning for a use of that expression in a given utterance context; where the assignment of this meaning-function to the relevant expression is generally a matter of (implicit or explicit) linguistic convention.

*Respective meaning ([sub-]propositional content, semantic content)*⁷ =_{DF} The respective meaning of an expression as used

in a given utterance context is a function yielding a reference or extension for that expression as used in that context, given particular circumstances of evaluation (see below).

In the case of indexical expressions, the respective meaning, alias semantic content, is a function of both the context of utterance and the general meaning-function of the expression used, which differs from the respective meaning; in all other cases, the two levels can be said to coincide.

Indexicality =_{DF} An expression is used as an indexical if and only if it is used in such a way that its respective meaning is dependent on both the utterance context (see below) and its general meaning-function, such that it may acquire different referents or extensions in different utterance contexts in virtue of its general meaning-function.

The level of respective meaning is subject to what Husserl calls “pure grammar,” which is the study of what distinguishes sense (i.e., respective meaning) from nonsense.⁸ On this view, semantic content displays something like formal, syntactic structure. This idea helps to explain the compositionality of meaning, which in turn explains how speakers and hearers, or interpreters, are able to grasp the meaning of an infinite number of sentences, many of which they have never heard before, on the basis of a finite vocabulary and a finite set of linguistic rules or conventions.

It is at the level of *respective* meaning that the bearers of truth-value (that is, of truth and falsity, respectively) are located—i.e., propositions. In modern semantics, truth-value ascriptions are relativized to what Kaplan calls *circumstances of evaluation*, consisting of possible worlds and, according to Kaplan, also times, on occasion. To illustrate one of the theoretical merits of this relativization to possible worlds, consider an utterance of mine of the sentence

⁶ The corresponding idea of different levels (*Stufen*) of understanding, which include the grasping of both character, content, and implicatures, is borrowed from Künne, who is also to be credited for pointing out the close similarity between Kaplan’s character/content distinction and Husserl’s distinction between general meaning-function and respective meaning; cf. Künne 1982. In Beyer 2000, I worked out the consequences of this distinction for Husserl’s semantics and theory of intentional content (“noematic sense”) in detail, arguing that the latter is to be rationally reconstructed as a moderate version of externalism, and that it can be fruitfully compared to Evans’ (radically externalist) neo-Fregean conception of sense, among others. That Husserl’s view can be read this way lends support to Dagfinn Føllesdal’s so-called Fregean interpretation of Husserl’s notion of noema (cf. Føllesdal 1969).

⁷ Note that “semantic content” is used by some authors to refer to conventional linguistic meaning rather than respective meaning (which Kaplan calls “content”).

⁸ Husserl’s investigations into pure grammar, especially his notion of a syntactic meaning category, had an important impact on modern linguistics (due mainly to Ajdukiewicz 1935).

(S0) I exist.

If we make the relativization in question, we can say two things: first, for every context of utterance it holds that the respective proposition expressed in an utterance of this sentence is true in the possible world of that context, so that the *sentence* can be said to be *a priori* or *logically true* (Kaplan 1989). Second, the *proposition* expressed in a particular utterance of S0, the respective meaning, is only *contingently true* – after all, the speaker need not exist: there are, in other words, possible worlds in which the proposition in question is false. Note that:

Context of utterance =_{DF} The utterance context consists of the possible world in which the utterance is (assumed to be) performed, the speaker, the addressee, the time and the place of utterance *and/or* all other entities that (according the general meaning-function of the expressions uttered) have to be identified in order to evaluate the utterance in terms of truth, falsity, or reference, relative to given circumstances of evaluation.

Or thus goes the rather common definition of “utterance context” I have used in earlier writings (e.g., Beyer 2001, pp. 278-279).

It is generally agreed upon, in mainstream semantics, that the levels of meaning mentioned so far – character and respective semantic content – do not exhaust what is communicated in speech. As Husserl puts it, there are mental states given voice to “in the broader sense,” and their contents are candidates for what the speaker non-literally means or suggests, which Grice calls “implicature.” At the same time, these contents are further candidates for what the hearer grasps when understanding, or successfully interpreting, the utterance.

This has been standardly regarded as a third level of meaning that is not the subject matter of formal semantics but rather of pragmatics: the study of the use of language for purposes of action other than the expression of literal meaning.

What is implicated (suggested, indirectly communicated) =_{DF} By using an expression in an utterance context, a speaker implicates the intentional contents of the mental acts she gives voice to in the broader sense. These contents can be made out on the basis of the respective meaning of the expression in that context by applying certain conversational maxims (cf. Maibauer 2010).

3 A contextualist challenge

This, then, is more or less the received opinion, which has been challenged by philosophers on the basis of ideas that partly go back to Husserl—in particular the notion of *background*. Thus, in his 1978 essay on “*Literal Meaning*” Searle claims that:

[...] for a large number of cases the notion of the literal meaning of a sentence only has application relative to a set of background assumptions, and furthermore these background assumptions are not all and could not all be realized in the semantic structure of the sentence in the way that presuppositions and indexically dependent elements of the sentence’s truth conditions are realized in the semantic structure of the sentence. (Searle 1978, p. 210)

On this view, the role of context is not simply that of fixing the reference of indexical expressions in a semantically well-regulated manner. There is contextual content determination everywhere, and correspondingly there is semantic underdetermination all over the place. There is no propositional meaning content attached to a sentence independently of context; and (some authors would add) *context* itself is not a well-defined notion: there is no neat list of semantically fixed context-factors and context-sensitive expressions. There is a huge and confusing background of assumptions, or know-how, that we bring to a given linguistic utterance, without which the utterance would fail to express any semantic content, and to thereby

determine truth conditions; and there is no hope of constructing a formal theory of this background (or “context”) and the way it determines truth-conditional content. Thus runs Searle’s radical contextualist challenge to mainstream semantics and pragmatics.

To motivate contextualism (so conceived) about meaning and content, consider a situation in which an object is hidden in a box. All we know about that object is that it is the only object in that box. Unlike us, the speaker knows which kind of object is in the box. She does not know that we do not know this; she intends to refer to a particular object of that kind, the one she takes to be in the box, or to one of its aspects (dependent features). She utters the sentence

(S1) This is red.

to make a statement about the object or aspect, without implying or suggesting anything else. What statement does she make? What is the respective meaning expressed in this utterance? What does the speaker say? According to radical contextualism, this depends on a wide variety of factors, not encoded in the linguistic meaning of the sentence uttered.

For a bird to be red (in the normal case), it should have most of the surface of its body red, though not its beak, legs, eyes, and of course its inner organs. Furthermore, the red color should be the bird’s natural color, since we normally regard a bird as being ‘really’ red even if it is painted white all over. A kitchen table, on the other hand, is red even if it is only painted red, and even if its ‘natural’ color underneath the paint is, say, white. Moreover, for a table to be red only its upper surface needs to be red, but not necessarily its legs and its bottom surface. Similarly, a red apple, as Quine pointed out, needs to be red only on the outside, but a red hat needs to be red only in its external upper surface, a red crystal is red both inside and outside, and a red watermelon is red only inside. [...] In short, what counts

for one type of thing to be red is not what counts for another. (Lahav 1989, p. 264)

So, in which way does the relevant meaning of S1 (“This is red”) depend on context? I want to consider three options.

1. *Speaker intentions*: Are the referential intentions of the speaker, such as their intention to refer to a particular *bird* by “this,” part of the relevant context? One problem with this answer is that it prevents us from adopting a conception of context according to which shared knowledge of context is what (in addition to shared knowledge of conventional linguistic meaning) enables both speaker and hearer to grasp one and the same respective meaning in cases of successful communication. After all, context, thus understood, is supposed to help the hearer make out the speaker’s referential intentions, among other things. So the present answer does not help—provided we conceive of context in a communication-theoretical way—as a means, so to speak, that in accordance with the relevant linguistic meaning enables the hearer to determine the respective meaning expressed.⁹
2. *Object referred to*: Is the relevant context simply identical to what’s in the box? But the speaker might only be referring to a particular aspect of the object in the box, rather than to the whole object. So we are thrown back to the speaker’s referential intentions—which do not help us, as we saw above.
3. *Background assumptions*: Does the relevant context consist of background assumptions about the object, or kind of object, in the box? Which assumptions, exactly? It seems to be impossible to make a comprehensive list, because every set of assumptions brings with it further assumptions. For example, suppose that the speaker takes an apple to be in the box. *Apples* normally count as red even if their *skin* is not completely red. However, consider a social group who have only encountered two kinds of apples thus far (as far as their colour

⁹ The epistemic availability of this means may require further means, to be found in a wider context itself not necessarily predelineated semantically.

is concerned): apples whose skin is completely red and apples whose skin is completely green; imagine that their apples instantaneously turn red when ripe. These people probably wouldn't classify an almost-ripe apple of the kind we know as "red." In fact, they wouldn't know what to say, because they have always assumed that there are only two kinds of apple-colour, and because this background assumption determines the meaning they conventionally associate with S1 as applied to apples. So shall we regard the assumption that there are grades of apple-redness corresponding to their ripeness as part of the context of our assertive uses of the sentence "this [the speaker refers to an apple] is red"? But how many grades are relevant? What if there had been exactly three apple colours? This would probably again lead to a different use, and hence respective meaning, of S1, as applied to apples, and so on and so forth.

Obviously these sorts of examples can easily be multiplied. Is there any way to avoid the following radical contextualist conclusion?

Radical contextualism =_{DF} There is no fixed relation between

- (i) the linguistic or literal meaning of a sentence S;
- (ii) a neatly defined set of context parameters; and
- (iii) the respective meaning and truth condition of S in the context of utterance, such that (iii) is uniquely determined by (i) and (ii).

Rather, the respective meaning is always determined differently, from situation to situation, so that the notion of a conventionally (co-)determined semantic content is untenable.

4 Two kinds of knowledge about truth conditions

The best strategy I can think of to avoid this radical conclusion draws upon a distinction made by Emma Borg. In her 2004 book *Minimal Semantics*, Borg distinguishes between minimal se-

mantic understanding, i.e., knowledge of what she calls "liberal" truth conditions, on the one hand, and knowing how to "verify" (or knowing what would make it the case) that the truth condition *is met*, on the other hand (Borg 2004, p. 238). Thus, the members of the social group who only know (what we would call) completely red-skinned and completely green-skinned apples are unable to know whether the truth condition of the sentence "this [the speaker refers to an apple] is red" is *met* regarding a not fully ripe apple, but they nevertheless *know* the truth condition—namely that the object the speaker wants them to attend to be red—whatever the latter may require in the case at hand. *They have full semantic knowledge but lack background know-how.* However, the latter is only required for "verification," or

1. knowledge of the proposition *p* stated (i.e., knowledge that *p*),

but not for the less demanding

2. knowledge of *which* proposition was stated (i.e., knowledge that *p* is the proposition literally expressed by the speaker).

The latter is sufficient for semantic knowledge regarding the statement.

I like this answer to the contextualist challenge, but I think that it eventually leads to a more moderate version of contextualism, rather than to a full-scale rejection: it leads to a version that makes room for semantic knowledge without background assumptions or know-how, knowledge whose content can indeed be investigated by formal semantics.

Clearly, the advocate of the present answer needs to explain how one can understand a sentence while lacking the kind of background know-how regarding which Searle would claim that in the absence of such capacities the "notion of the meaning of the sentence" has no clear "application" at all (see quotation above). Searle would stress that in the absence of appropriate background assumptions or know-how we have no clear idea of how to understand a sentence like "This (apple) is red;" which mani-

festes itself in the fact that we do not, for instance, know how to follow the corresponding order “Bring me the red apple!” (cf. [Searle 1983](#), p. 147). In the light of Borg’s distinction, this can be described as lack of knowledge about “verification,” but what about the strong intuition that in the absence of such background know-how the sentence fails to express a content that can be evaluated in terms of truth and falsity? To strengthen this intuition, consider Searle’s examples S2–S4 (cf. [Searle 1983](#), Ch. 6):

- (S2) Bill opened the mountain.
- (S3) Bill opened the grass.
- (S4) Bill opened the sun.

These sentences are syntactically well-formed and contain meaningful English expressions; yet they do not express clear semantic content—unless we imagine some background know-how regarding what it means to open a mountain, the grass, or the sun.¹⁰ The mere combination of the literal meaning of the verb “opened” with the literal meanings of other English expressions in accordance with the English syntax does not seem to be enough to produce a clear truth-evaluable content, despite the fact that “to open” does not look like an indexical that yields as reference a unique behavioural relation (or type of action) referred to as “opening,” for a neatly defined type of context—in the way that “I” always yields the speaker of the utterance context as its referent. Borg would disagree; she says about an analogous example by Searle (“John cut the sun”):

If the competent language user understands all the parts of the sentence (she knows the property denoted by the term ‘cut’, she grasps the meaning of the referring term ‘John’ and she understands the meaning of the definite description ‘the sun’) and she understands this construction of parts, then she knows that the utterance of this sentence is true just in case

[...] John stands in the cutting relation to the sun. Now clearly any world which satisfies this condition is going to be pretty unusual (and there may be some vague cases [...]) but there will be, it seems, some pretty clear cases on either side of the divide. For instance, any world where John’s actions do not have any effect on the physical status of the sun is clearly going to be a world where the truth-condition is not satisfied. While any world where John’s actions do result in some kind of severing of the physical unity of the mass of the sun is a world where the truth-condition is satisfied. ([Borg 2004](#), p. 236)

This reply to Searle is unconvincing for at least two reasons.

First: To begin with, Borg here equates semantic knowledge concerning the verb phrase “cut” with knowledge of the property it denotes (see the first brackets in the quotation). But arguably this phrase does not denote any property in isolation; it only does so in the context of a *sentence* (by the “context principle”).¹¹ And Searle’s parallel point about “opened” is that this verb phrase denotes quite *different* properties in S2–S4, respectively, without being ambiguous. That the verb is unambiguous in these cases becomes intuitively plausible if we apply the “conjunction reduction” test (cf. [Searle 1992](#), pp. 178–179). Instead of asserting the conjunction of S2–S4 we can just as well say: “Bill opened the mountain, the grass, and the sun” and perhaps add: “he used a secret universal device for the task recently developed by NASA.” This may be a weird example, but its

¹¹ Cf. [Beyer 1997](#), p. 341, where I raise the same point in order to criticize one of Searle’s arguments for the Background Hypothesis. As for the precise content of the context principle, Robert Stainton distinguishes between three readings:

“The first [is] merely methodological, a claim about how to find out what particular words mean: To find word meanings, look at what they contribute to sentences. The second reading [is] metasegmental, a claim about why words have the meanings they do: words only have meaning because of how they affect sentence meanings. The third reading of the Principle is interpretational/psychological. [...] [T]he idea underlying it is that the only things we are psychologically able to understand are whole sentences.” ([Stainton 2010](#), pp. 88–89). In the present context, a consequence of the *metasegmental* reading is intended which follows from the conjunction of that reading and the assumption that the meaning of a predicate (like “... cut ...”) denotes a property or relation, if anything.

¹⁰ Another option might be to admit category mistakes as semantic contents. (I wish to express my thanks to Adriana Pavic for reminding me of this option.)

weirdness does not seem to be due to the ambiguity of “opened.” Rather, unlike the imagined NASA devisors we simply have no background know-how that would enable us to assign truth conditions to this sentence.

Borg would probably reject the context principle (thus paying a high price for her view) and answer that there may be vague cases in which we do not know whether the opening relation obtains or not, but that “there will be [...] some pretty clear cases on either side of the divide” (Borg 2004, p. 236); after all, in the preceding quotation she makes a parallel claim about the example “John cut the sun.” But this answer is, again, unconvincing (as is Borg’s parallel claim). One might just as well argue that both S5 and S6 describe the same relation, the opening relation, as obtaining between different objects.

(S5) Bill opened his hand.

(S6) Bill opened the door.

But opening a hand is an intentional bodily movement, while opening a door is a more advanced or complex action that merely *involves* such bodily movements. These are different kinds of behavioural relation. Of course, clear examples of the obtaining of both of these relations may have something in common, but this common feature does not seem to constitute a common *type of action*. And what (if anything) is the verb phrase in S5 and S6 supposed to denote, if not a type of action?

Nor is the verb phrase in this pair of sentences ambiguous. This is made plausible by the conjunction reduction test: it is perfectly fine to abbreviate the conjunction of S5 and S6 as follows: “Bill opened the hand and the door.”

The (to my mind) false impression that the unambiguous verb phrase in S2–S4 denotes the same behavioural relation or feature as in, say, S6, merely comes from the fact that we tend to think of *established* uses of “*a* opened *b*” sentences (or “*a* cut *b*” sentences) when trying to construct an interpretation for cases like S2–S4 that we do not really understand. But there is no such use in these cases (see the next paragraph but one).

Second: Moreover, Borg’s claim that “any world where John’s actions do result in some kind of severing of the physical unity of the mass of the sun is a world where the truth-condition [of ‘John cut the sun’] is satisfied” is simply false. If John causes an explosion whose effect is that the physical unity of the mass of the sun is severed (such that it breaks into, say, two halves),¹² he does not thereby *cut* the sun. I suppose that any attempt to secure a minimal truth condition for S4 (and S2–S3, for that matter) is doomed to failure. In order to have at least a slight chance of getting off the ground, any such attempt will have to mention something that can be done using sharp-edged tools (or devices simulating such tools),¹³ and it seems impossible to define (let alone imagine) a procedure of this type that could in principle be applied to the sun.

To anticipate the alternative approach I am going to take, in cases like S2–S4 there is no established sentence-use because there is no appropriate background know-how to be found in the relevant social group (including its late members), hence no group of (current or former) “producers” (see below), and hence no relation conventionally denoted by the verb phrase that could enter the respective truth condition. Therefore, these sentences have “literal meaning” (as Searle puts it) but lack semantic content. Literal meaning is not usage (in the current sense), nor does it require a particular usage—unlike respective meaning.

On similar grounds (to return to the last example about S1), if in the envisaged social group there is no background know-how regarding certain apples that *we* would readily classify as “red,” against that background, the sentence S1 has no clear application to such apples in the language use of that group, and it *does not have the same truth condition* as in ours. An in-

¹² A reviewer claims that “to sever” means to cut. Even if the corresponding interpretation of “severing” were admissible, it could not be the one intended by Borg. Have a look at the preceding quotation. If you replace “severing” by “cutting” there, you obtain: “While any world where John’s actions do result in some kind of cutting of the physical unity of the mass of the sun is a world where the truth-condition [of ‘John cut the sun’] is satisfied.” If this sentence is meaningful at all, it expresses a triviality that does nothing to support Borg’s view.

¹³ See the entry on “cut” in the Oxford Advanced Learner’s Dictionary of Current English.

interpretation problem occurs. I am attracted by an interpretation-theoretical principle proposed by Timothy Williamson, probably inspired by Gareth Evans, which Williamson calls the principle of knowledge maximization (as opposed to the principle of *truth* maximization to be found in traditional hermeneutics, and endorsed by Donald Davidson):

The shift from conventions of truthfulness to conventions of knowledgeableness also has repercussions in the methodology of interpretation. The appropriate principle of charity will give high marks to interpretations on which speakers tend to assert what they know, rather than to those on which they tend to assert what is true [...]. (Williamson 2000, p. 267)

The right charitable injunction for an assignment of reference is to maximize knowledge, not to minimize ignorance. (Williamson 2007, p. 265)

According to the principle of knowledge maximization, an interpretation is correct to the extent that it maximizes “the number of knowledgeable judgements, both verbalized and un-verbalized, the speaker comes out at making” (McGlynn 2012, p. 392). To motivate this principle, although in a somewhat modified version, imagine that in the above example about the box there are in fact two objects in the box—a red ball and a yellow apple—but that we know that the speaker does not know about the ball, which was already hidden in the box before we put the apple into the box while the lightning was such as to make the apple look red.¹⁴ The speaker, who observed how we put the apple into the box, mistakenly believes it to be red and exclaims: S1 (“this is red”). No doubt, if this utterance has any truth condition, it involves an apple rather than a ball, and the utterance is false. A suitably modified version of

the principle of knowledge maximization yields this result as the correct interpretation, while the principle of truth maximization fails to do so. After all, the speaker would only give voice to a true belief here if her statement concerned the *ball* rather than the apple. However, this belief would not qualify as knowledge, in the described situation, and by assumption the speaker lacks any other knowledge regarding that ball. By contrast, the speaker possesses some knowledge about the apple, which is in fact yellow. In Evans’ terms, the speaker has opened a mental dossier (a dynamic system of beliefs) about the apple, which contains quite a number of (correct) information about it, even though the addition of the belief that the apple is red does not enlarge that body of knowledge. Thus, the speaker ought to be interpreted as giving voice to that false belief, Davidson and traditional hermeneutics notwithstanding. This interpretation takes into account more relevant knowledge on the part of the speaker than the other.

The principle of knowledge maximization needs to be modified in terms of, or supplemented by, a more traditional theory of justification in order to yield this result. To see this, let us first consider another example, inspired by Husserl (cf. Husserl 1987, p. 212), which I have used in earlier writings to motivate my “neo-Husserlian,” moderately externalist reconstruction of his view on respective meaning and intentional content.

Let’s assume that at a time t_1 Ed points at a certain table in the seminar room where he has just been lecturing and exclaims:

[(S7)] This table wobbles.

One of the students is prepared to take Ed to the caretaker, to make sure that the table gets repaired immediately. The way from the seminar room to the caretaker’s office is rather complicated. But they manage to find it. The caretaker asks Ed to take him to the seminar room with the wobbling table. The student has other things to do. So Ed has to take the care-

¹⁴ Following the realism inherent to ordinary language use, I assume that the everyday world of experience involves objects displaying real colours. It may be possible to eliminate real colours, but such attempts at revisionary metaphysics should have no impact on the study of the actual use of language, unless they lead to a change of language use, which has not happened yet in the case of colour words.

taker to that room by himself. Finally, they arrive at a seminar room that Ed falsely believes to be the room with the wobbling table. At t_2 Ed points at a certain table, which he regards as that wobbling table, and once again declares [S7]. The caretaker investigates the table and contradicts Ed—who reacts somewhat irritatedly. (Beyer 2001, pp. 284–285)

It is unclear which referent (table 1 or table 2) the interpreter is supposed to assign to the demonstrative term “this” according to the (unmodified) principle of knowledge maximization. After all, both of these assignments would lead to an ascription of knowledge to the speaker: knowledge about table 1 (to the effect that it wobbles) and table 2 (to the effect that it is a table he takes to be wobbling), respectively.

To decide the issue, the interpreter needs to take a closer look at the speaker’s epistemic motivation for making the judgment given voice to in her utterance of S7 at t_2 —he needs to consider the experience(s) with recourse to which the speaker can *justify* her claim to knowledge. If the judgment is motivated by a perception the speaker is having, thus qualifying as an *observational* judgment, it will be about the object of that perception: that object is perceived as thus-and-so and for this reason (on this ground) judged to be thus-and-so. This is what happens at t_1 : the speaker perceives table 1 as wobbling and is sincerely giving voice (in the narrow sense) to an accordingly motivated judgment to the effect that it wobbles. However, at t_2 the epistemic situation is different. The speaker’s judgment is motivated by a *memory of table 1* rather than by her current perception of table 2. It is this memory that rationalizes her judgment, and could be self-ascribed by the speaker when justifying her judgment. Therefore, the speaker gives voice, in the narrow sense, to a judgment about table 1, namely that it wobbles. I have elsewhere called this epistemically-determined truth condition the utterance’s *internal truth condition*.¹⁵ According to

the neo-Husserlian approach, respective utterance meaning determines the *internal* truth condition.

So in order to yield interpretations that adequately reflect the meaning intentions actually given voice to by the speaker, and thus the respective meanings of their utterances, the principle of knowledge maximization needs to be supplemented by (or reformulated in terms of) a more traditional theory of justification, drawing upon notions like *observation*, *memory* and *testimony* (referring to sources of justification). Note that the present approach to reference supports a version of the context principle: it is only in the context of a judgment that a referent can be assigned to a mental act of reference given voice to by a singular term.

Let us finally return to the example about the two objects in the box. In this example the speaker gives voice to a judgment about the yellow apple rather than the red ball in her utterance of S1, because she (falsely) remembers that ball as being red, having opened, on an earlier occasion, a mental dossier about it containing the (incorrect) information that the ball is red, while she neither remembers nor perceives, nor has heard about the ball that also happens to be in the box. Thus, the judgment given voice to can only be motivated by, and justified with recourse to, that memory—even if it does not yield knowledge in the case at hand. And that memory concerns the apple rather than the ball, because it belongs to the speaker’s body of information about the apple. Thus, on a version of the principle of knowledge maximization modified in accordance with the foregoing neo-Husserlian approach to reference assignment, the utterance in question concerns the apple, if anything.

Now by the principle of knowledge maximization (in both versions), if there is no back-

ternal truth-condition is the state of affairs represented by the (intentional content of the) judgement the speaker should give voice to, given (a) the linguistic meaning [i.e., the general meaning-function] of the employed sentence and (b) the external context.” The external context is the actual (observable) context of utterance, which on the neo-Husserlian approach may differ from the phenomenologically relevant (“internal”) context, which is determined by the motivational structure of experience with recourse to which the speaker could justify the judgment given voice to. In the present example, the internal context involves table 1.

¹⁵ Cf. Beyer 2001, p. 289: “The internal truth-condition of an assertion is the state of affairs represented by the (intentional content of the) judgement actually given voice to in that assertion. Whereas the ex-

ground that enables members of a social group to express knowledge by a sentence like S1 in a situation where we would readily apply that sentence—on the basis of our own knowledge *and* background know-how—, then the following conclusion recommends itself: in the language use of this social group the sentence lacks the determinate truth-evaluable meaning it expresses in our own language use, in a given context. (Contrast what Borg says about “John cut the sun;” see the above quotation from [Borg 2004](#), p. 236.)

This speaks in favour of contextualism. However, it does not speak in favour of a radical version of contextualism, which would not allow for a notion of minimal semantic knowledge that can indeed be possessed in the absence of personal background know-how—a version that thus ignores the above-described difference between two types of knowledge regarding truth conditions. In what follows, I shall sketch a more moderate version of contextualism that does take this difference into account.

5 Towards a moderate contextualism about meaning

It is plausible to assume that all that is required for semantic knowledge, conceived as knowledge *which* truth condition has been stated, is that the following two conditions be met. First, a sufficient number of current or former (late) members of the *speech community* to which the speaker belongs possess appropriate background know-how. Second, the speaker stands in an appropriate social relation to these members (a relation that would enable the speaker to express communal knowledge by a true sentence whose content she may be unable to “verify” herself).

These members are experts; they are capable of “verifying” or “falsifying” the semantic content of the sentence in question, as opposed to merely grasping it. Other members of their social group participate in their knowledge thanks to intersubjective processes of information transfer. The main idea behind this approach is an adaptation of Evans’ distinction between what he calls name-producers and mere name-consumers, which is used to substantiate

the above distinction between two kinds of knowledge regarding a sentence’s truth condition.¹⁶ This strategy leads to a social-epistemological conception of the background of meaning and to a version of contextualism that preserves basic insights of anti-contextualists like Borg. Evans writes:

Let us consider an ordinary proper-name-using practice, in which the name ‘NN’ is used to refer to the person x. The distinctive mark of any such practice is the existence of a core group of speakers who have been introduced to the practice via their acquaintance with x. They have on some occasion been told, or anyway have come to learn, a truth which they could then express as ‘This is NN’, where ‘This’ makes a demonstrative reference to x. Once a speaker has learned such a truth, the capacity to re-identify persons over time enables him to recognize later occasions on which the judgement ‘This is NN’ may be made, and hence in connection with which the name ‘NN’ may be used. [...] Members of this core group, whom I shall call ‘producers’ [...], do more than merely use the name to refer to x; they have dealings with x from time to time, and use the name in those dealings – they know x, and further, they know x as NN. [...] [T]he expression does not become a name for x unless it has a certain currency among those who know x – only then can we say that x is known as NN. [...] Perhaps in the early stages of its existence all the participants in the name-using practice will be producers, but this is unlikely to remain so for

¹⁶ As Evans acknowledges, this distinction is inspired by Putnam’s notion of a “linguistic division of labour” (see [Putnam 1975](#), pp. 145–146); cf. [Evans 1982](#), p. 377. I should stress that on the view proposed in this contribution, the producers do not grasp the respective meaning of relevant expressions more “fully” than the mere consumers. Rather, they help sustain the common practice necessary for those expressions to be usable (by both producers and mere consumers) to express a respective meaning (a truth-conditional “semantic content”). I should also stress that I take the producer/consumer distinction to be universally applicable, and not just in the case of rigid designators, and that on my view the capacities of the producers (unlike the capacities of what Putnam calls “experts”) need not include scientific knowledge. (Thanks to Adriana Pavic for pressing me on these points.)

long. Others, who are not acquainted with x , can be introduced into the practice, either by helpful explanations of the form ‘NN is the φ ’, or just by hearing sentences in which the name is used. I shall call these members ‘consumers’, since on the whole they are not able to inject new information into the practice, but must rely upon the information-gathering transactions of the producers. [...] Let us now consider the last phase of a practice of a name-using practice, when all the participants are consumers. [...]. Later consumers manifest the intention to be participating in this practice, and, using a name which, in the practice, refers to Livingstone, themselves refer to Livingstone. Thus the practice is maintained with a constant reference, perhaps for very long periods of time. (Evans 1982, pp. 376-393)

is satisfied; they know how to follow a corresponding order, and so on.

The rest of the speech community merely knows the truth condition and can gain and transfer information an utterance of the sentence bears without themselves being in the know—that is, without having the original knowledge only the producers have in their possession. They may acquire and transfer knowledge (sometimes) by testimony, thanks to the existence of a community-wide practice of sentence-usage sustained by intersubjective processes of information transfer, in a way yet to be understood in more detail.

Eventually, mere consumers “must rely upon the information-gathering transactions of the producers,” to use Evans’ formulation. Mere consumers have semantic knowledge, but they lack more substantive knowledge. Semantics is concerned with the content of their semantic knowledge. Mere consumers need a background of what Searle calls *social practices*, including social practices of language use. However, they lack the producers’ individual or personal background know-how and thus their substantive knowledge regarding truth conditions, which requires such know-how (i.e., the knowledge of how to “verify” those conditions).

What kind of individual background do the producers need in order to be able to make possible social practices of language use that allow all members of their speech community to express and grasp semantic contents determining particular truth conditions? In his 1978 paper, which some regard as the constitutive document of contextualism, Searle stresses the importance of background assumptions, such as the assumption that there is a field of gravitation or that things offer resistance to pressure, which is usually taken for granted, quite unreflectedly, when we speak about middle-sized everyday objects such as apples and boxes. This may at first sound like the requirement of what might be called background *knowledge*, consisting of intentional states, i.e., certain epistemically distinguished beliefs. However, especially in his later writings, Searle stresses the non-intentional character of the background, characterizing it as consisting of non-intentional capacities

If we adapt Evans’ distinction between two types of name-users for present purposes, we can say that in a given community there have to be, or have to have been (see the last three sentences of the quotation), people “in the know” regarding (what we use to call) the red colour of apples, or regarding a particular practice of opening mountains, grass, or the sun, in order for the sentences S2–S4 to be candidates, in virtue of their literal meaning, for the expression of knowledge available to us through these sentences.¹⁷ There have to be (current or late) “producers” in order for these sentences to express a semantic content determining truth conditions, thus displaying a clear, interpretable respective meaning in that linguistic community—and this requires that the sentences have a community-wide usage upheld by recourse to (current or late) producers. They know (or knew) how to “verify” the respective meaning of assertive utterances of the sentence, in the relevant usage; i.e., they know which fact (if any) would make it the case that the truth condition

¹⁷ The point is not that we cannot describe uncommon practices (such as using a metal saw) for actions like opening a can, say. Rather, the point is that there have to be common practices, known to the producers, in order for a sentence like “Bill opened the can” to be usable to express a respective meaning representing any practice in the first place. (Thanks again to Adriana Pavic for helping me to make this clear.)

—which I have referred to above as background *know-how*, and which would include the ability to perform social practices. Searle has formulated a thesis about the relation between intentionality, on the one hand, and background know-how on the other, a thesis he calls the “hypothesis of the Background”:

Another way to state [the hypothesis of the Background] is to say that all representation, whether in language, thought, or experience, only succeeds given a set of nonrepresentational capacities. In my technical jargon, intentional states only determine conditions of satisfaction relative to a set of capacities that are not themselves intentional. (Searle 1992, p. 175)

Later in the same book chapter he explains:

The actual content [sc. of an intentional state] is insufficient to determine the conditions of satisfaction. [...] Even if you spell out all contents of the mind as a set of conscious rules, thoughts, beliefs, etc., you still require a set of Background capacities for their interpretation. (Searle 1992, pp. 189-190)

This addition to the formulation in the penultimate quotation is, I think, false—or even absurd. The respective meaning of an utterance is the intentional content of the mental state given voice to in the narrow sense, which means that intentional content is precisely what determines the truth condition (or, more generally, the conditions of satisfaction). Indeed, Searle seems to agree:

[...] I want to capture our ordinary intuition that the man who has the belief that Sally cut the cake has a belief with exactly the same propositional content as the literal assertion ‘Sally cut the cake.’ (Searle 1992, p. 184)

I take it to be a definitional truth that intentional content provides the answer to Wittgenstein’s question “What makes my representation

of him a representation of *him?*”. A conception of intentional content must spell out this answer. It makes no sense to conceive intentional content along the lines of Searle’s addition in the penultimate quotation, just as it makes no sense (*pace* Searle) to say of semantic content, properly construed, that it is not self-applying, or that it needs to be interpreted against a non-representational background in order to determine reference or satisfaction conditions.

In the following passage Searle commits himself to radical contextualism:

An utterance of [the sentence ‘Sally gave John the key, and he opened the door’] would normally convey that first Sally gave John the key, and later he opened the door, and that he opened the door with the key. There is much discussion about the mechanisms by which this additional content is conveyed, given that it is not encoded in the literal meaning of the sentence. The suggestion, surely correct, is that sentence meaning, at least to a certain extent, underdetermines what the speaker says when he utters the sentence. Now, the claim I’m making is: sentence meaning radically underdetermines the content of what is said. (Searle 1992, p. 181)

Thus, Searle explains, nothing in the literal meaning of the sentence referred to excludes crazy interpretations like: “John opened the door with the key by swallowing both door and key, and moving the key into the lock by way of the peristaltic contraction of his gut.” (Searle 1992, p. 182) Note that we are dealing with a claim about linguistic meaning here, not about semantic content—properly construed as *representational* content, uniquely determining satisfaction conditions.

From the viewpoint of the social-epistemological picture of semantic content sketched above, the Background Hypothesis should be restricted to the *producers* of sentences figuring in linguistic representation. On this picture, only the producers’ intentionality requires background know-how regarding the application of

those sentences. Mere consumers merely need an appropriate background of *social practices*. If the advocate of this picture did not restrict the Background Hypothesis to the producers, he would be committed to the view that mere consumers can give voice, in the narrow sense, to intentional states in which they cannot be, due to lack of background know-how. This would mean that only the producers can be *sincere* in their assertive utterances of sentences regarding which they are producers. But this seems wrong. It is possible for mere consumers to deliberately express knowledge by testimony. Hence, the (unrestricted) hypothesis of the Background ought to be rejected, on the present view. Meaning-intentions (meaning-bestowing acts) do not generally require a non-intentional background relative to which their (truth-conditional) content and satisfaction conditions are determined; while their intuitive fulfilments (the corresponding “verifications”), if any, do. For instance, it is impossible to *perceive* something as an elm without being able to distinguish elms from other sorts of trees. This of course means in turn that one ought to reject the present picture if one accepts the Background Hypothesis (in unrestricted form). In order to decide the issue, more needs to be said to explain this hypothesis. I cannot decide the issue here. But it may be helpful in this regard to end by saying a bit more about the content of Searle’s Background Hypothesis.

In *The Rediscovery of the Mind* Searle plausibly contends that *mental representation*, i.e., underived, original intentionality is realized just in case a given mental state “is at least potentially conscious” (Searle 1992, p. 132). We find similar claims in Husserl.¹⁸ Due to the “as-

pectual shape” of intentional states (the fact that they have perspectival, intentional content) there are no “deep unconscious mental intentional phenomena” (Searle 1992, p. 173), such as reflectively inaccessible belief states. There is an important sort of background elements whose distinctive mark is that they are *capacities to be in intentional states*; that is, they are dispositions to have (actually or potentially) conscious representations, such as occurrent beliefs. The general assumption that things offer resistance to pressure is a case in point. We normally do not form a belief to this effect but are nevertheless committed to it by the way we behave towards things (cf. Searle 1992, p. 185).

One may call these capacities for (at least potentially) conscious representation “background assumptions” or “network beliefs” if one likes, but according to Searle one must keep in mind that these capacities fail to be intentional states: “the Network of unconscious intentionality is part of the Background” (Searle 1992, p. 188) and “the Background is not itself intentional” (Searle 1992, p. 196). If Searle is right about this, then many elements of the so-called “web of belief” are part of the non-intentional background.

This view has far-reaching consequences for the theory of intentionality. For, if Husserl is basically right about the structure of consciousness (as I believe he is), then conscious states must be embedded in a holistic structure, which Husserl calls the “intentional horizon,” whose future elements are predelineated (at least in part) by the intentional content of the respective state of consciousness. For example, if you consciously see something whose front side you

consciousness iff they are both intentional objects of a dispositional higher-order belief of the sort “I am now having such-and-such experiences” that would be actualized by one and the same higher-order judgment (where the temporal demonstrative specifically refers to the moment of (internal) time at which both of these experiences occur). (2) Two diachronous intentional experiences belong to the same stream of consciousness iff both of them are intentional objects of a dispositional higher-order belief of the sort “I just (or earlier) had such-and-such experiences” that would be actualized by one and the same higher-order judgment. This approach fits in well with Husserl’s contention that “[i]ntentionality is what [...] justifies designating the whole stream of [experiences] as the stream of consciousness and as the unity of *one* consciousness” (Husserl 1982, p. 199). It also fits in well with a view on which Husserl conceives of consciousness as “pre-reflective self-awareness;” cf. Beyer 2011.

¹⁸ Husserl has a dispositionalist higher-order judgment view of consciousness, according to which conscious experiences are “essentially capable of being perceived in reflection,” such that “they are there already as a ‘background’ when they are not reflected on and thus of essential necessity are ‘ready to be perceived’” (Husserl 1982, p. 99; also cf. p. 80, where Husserl cites as an example a case in which “we are reflecting on a conviction which is alive right now (perhaps stating: I am convinced that ...)”). (Compare Searle 1992, p. 156: “This idea, that all unconscious intentional states are in principle accessible to consciousness, I call the connection principle [...].”) In Beyer 2006, Chs. 1-2, I defend a dispositionalist higher-order judgment view of intentional consciousness and argue that it explains the unity of consciousness (1) at a time as well as (2) across time, as follows: (1) Two simultaneous intentional experiences belong to the same stream of con-

are visually confronted by *as a house*, then you will anticipate¹⁹ visual appearances of a back side and an inside, respectively, as future experiences you would undergo if you walked inside or walked around the object while observing it. But is the corresponding set of anticipations really an *intentional* structure? Searle's arguments regarding the background cast doubt on this, given his view that consciousness (or what is consciously accessible) is the only occurrent reality of intentionality. After all, it is plausible to equate (a large subset of) the set of anticipations determining the respective intentional horizon with a relevant part of what Searle calls the "Network," given that they cannot be described properly as occurrent beliefs or conscious judgments, but rather as mere dispositions to form higher-order beliefs. For, as Husserl explains, the anticipations in question concern the way the represented object would present itself to consciousness in possible worlds compatible with what is currently experienced, and they also concern the way this object relates to other objects in the world—thus constituting the core of one's current *world* horizon, which core Husserl calls the "external horizon" (Husserl 1973, p. 32) of the experience (see below). It is only when these anticipations are intuitively fulfilled, in the sense that relevant conscious episodes of (what seem like) *verification* (such as perceptual verification) occur, motivating corresponding acts of judgment, that there will be entries into the relevant mental dossier associated with the object in question. As Husserl puts it (referring to mental dossiers associated with proper names as "individual notions"):

I see an *object without an 'historic' horizon* [footnote: without a horizon of acquaintance and knowledge], and now it gets one. I have experienced the object multifariously, I have made 'multifarious' judgements about it and have gained multifarious [pieces of] knowledge about it, at various times, all of which I have connec-

ted. Thanks to this connection I now possess a 'notion' of the object, an individual notion [...] [W]hat is posited in memory under a certain sense gains an epistemic enrichment of sense, i.e., the *x* of the sense is determined further in an empirical way.²⁰ (Husserl 2005, p. 358; my translation)

The "historic" horizon and the objects of the relevant anticipations constitute the "internal horizon" (Husserl 1973, p. 32) of the experience. They all belong to the same "*x* of the sense" (also referred to by Husserl as the "determinable *X*"), i.e., they share a sense of identity (of represented object) through time. Other past and anticipated experiences bring it about that one's "'notion' of the object" is *networked* with other notions of objects. They constitute the external horizon of the experience.

If the anticipations in question were part of a non-intentional background, then it would be wrong, of course, to describe them as being directed at objects; as a consequence, the Husserlian conception of intentional horizon just sketched would break down. To avoid this consequence, Searle's Background conception needs to be altered, such that the background may indeed contain intentional elements, albeit in a derived sense.

This can be fleshed out as follows. The primary bearers of intentionality are (at least potentially) conscious units, such as judgments and the experiences that motivate them. It is true that respective meaning and intentional content only function against a background the elements of which lack this primary form of intentionality. However, this background contains some elements that possess a *derived* form of intentionality, so that it is misleading to describe it as completely non-intentional. In particular,

²⁰ The German original runs: „Ich sehe einen *Gegenstand ohne einen „historischen“ Horizont* [Fn.: ohne Bekanntheithorizont und Wissen-shorizont], und nun bekommt er ihn. Ich habe den Gegenstand vielfältig erfahren, „vielfältige“ Urteile habe ich über ihn gefällt, vielfältige Kenntnis von ihm in verschiedenen Zeiten gewonnen und habe sie verknüpft. Nun habe ich durch diese Verknüpfung einen „Begriff“ von dem Gegenstand, einen Eigenbegriff [...]. [D]as in [der Erinnerung] mit einem gewissen Sinn Gesetzte erfährt eine erkennt-nismäßige Sinnbereicherung, das heißt, das *x* des Sinnes bestimmt sich näher erfahrungsmäßig.“ (Husserl 2005, p. 358).

¹⁹ For the close connection between anticipation and (internal) horizon, cf. Husserl 1973, para. 8. For an insightful interpretation of Husserl's notion of horizon, cf. Smith & McIntyre 1982, pp. 227–265.

it contains mental capacities or dispositions to form beliefs about the further course of experience which Husserl (in 1973, para. 8) calls “anticipations.” Some of the experiences thus anticipated correlate with an internal horizon. Their occurrence may lead to entries being made in a mental dossier, which are empirical beliefs (informational states) to which a “referent” (an object they are about) can be assigned in a principled way, in accordance with the modified principle of knowledge maximization. Here is an example of such a principle of reference assignment, which I have proposed in earlier work.²¹

The logical subject x of [...] a belief of the form a is F [...] whose acquisition goes together with the opening of a mental dossier about x is identical with the logical subject y of the judgement *initiating* that belief (or x would be identical with y , if x and y existed). (Beyer 2001, p. 287)

“Logical subject” here refers to the object the relevant belief is about (such as table 1 in the case of the persisting belief actualized by the judgment given voice to at t_2 in the above example about the wobbling table); and the judgment initiating that belief is understood to have its logical subject assigned in accordance with the modified principle of knowledge maximization, as explained at the end of section 3, above.

I conclude, first, that the background of meaning and intentional content may be looked upon as being at least in part itself intentional, albeit in a derived sense, but that, second, the applicability of the Background Hypothesis still needs to be restricted, as far as the part of the background (co-)determining truth-conditional content is concerned, to what I have called the producers.

6 Conclusion

In summary, I have distinguished three levels of meaning, the first of which (general meaning-function) is a matter of linguistic convention, while the second level (respective meaning) is

truth-conditional and partly dependent on the first, purely semantic level, but also dependent on the reference or extension determined by the intentional state actually given voice to. This intentional state has its intentional object (the reference of the corresponding utterance) fixed epistemically, in accordance with the modified principle of knowledge maximization. Furthermore, this epistemic reference-fixing depends on the informational states (or dossiers) of the producers only. Only the producers need to possess the kind of background that Searle wrongly takes to be required for all speakers or hearers capable of giving voice to or grasping the respective meaning in question, the grasping of which then serves as the basis for accessing the third, purely pragmatic level of meaning (namely, what is implicated).

Acknowledgements

Earlier versions of this paper were presented to audiences in Göttingen and Erfurt, whom I thank for helpful discussion. For helpful comments and suggestions I would like to thank the editors and three anonymous reviewers. I am particularly grateful to Adriana Pavic for her detailed comments and suggestions.

²¹ For further neo-Husserlian principles of reference assignment, see Beyer 2000, para. 7; Beyer 2001.

References

- Ajdukiewicz, K. (1935). Die syntaktische Konnexität. *Studia Philosophica*, 1, 1-27.
- Barber, A. & Stainton, R. (Eds.) (2010). *Concise encyclopedia of philosophy of language and linguistics*. Amsterdam, NL: Elsevier.
- Beyer, C. (1997). Husserle's representationalism and the "hypothesis of the Background". *Synthese*, 112 (3), 323-352. [10.1023/A:1004992424269](https://doi.org/10.1023/A:1004992424269)
- (2000). *Intentionalität und Referenz*. Paderborn, GER: mentis Verlag.
- (2001). A neo-Husserlian theory of speaker's reference. *Erkenntnis*, 54 (3), 277-297. [10.1023/A:1010795215502](https://doi.org/10.1023/A:1010795215502)
- (2006). *Subjektivität, Intersubjektivität, Personalität*. Berlin, GER: De Gruyter.
- (2011). Husserls Konzeption des Bewusstseins. In K. Cramer & C. Beyer (Eds.) *Edmund Husserl 1859-2009* (pp. 43-54). Berlin, GER: De Gruyter.
- Beyer, C. & Weichold, M. (2011). Philosophy of language. In S. Luft & S. Overgaard (Eds.) *The Routledge companion to phenomenology* (pp. 406-416). London, UK: Routledge.
- Borg, E. (2004). *Minimal semantics*. Oxford, UK: Oxford University Press.
- Evans, G. (1982). *The varieties of reference*. Oxford, UK: Clarendon Press.
- Føllesdal, D. (1969). Husserl's notion of noema. *The Journal of Philosophy*, 66 (20), 680-688. [10.2307/2024451](https://doi.org/10.2307/2024451)
- Husserl, E. (1973). *Experience and judgment*. Evanston, IL: Northwestern University Press.
- (1982). *Ideas pertaining to a pure phenomenology and to a phenomenological philosophy*. Den Haag, NL: Nijhoff.
- (1987). *Vorlesungen über Bedeutungslehre Sommersemester 1908*. Dordrecht, NL: Nijhoff.
- (2001). *Logical investigations, vol. 1*. London, UK: Routledge.
- (2005). *Logische Untersuchungen Ergänzungsband, Zweiter Teil: Texte für die Neufassung der VI. Untersuchung: Zur Phänomenologie des Ausdrucks und der Erkenntnis (1893/94-1921)*. Dordrecht, NL: Springer.
- Kaplan, D. (1989). Demonstratives. In J. Almog, J. Perry & H. Wettstein (Eds.) *Themes from Kaplan* (pp. 481-563). New York, UK: Oxford University Press.
- Künne, W. (1982). Indexikalität, Sinn und propositionaler Gehalt. *Grazer Philosophische Studien*, 18, 41-74. [10.5840/gps1982183](https://doi.org/10.5840/gps1982183)
- Lahav, R. (1989). Against compositionality: The case of adjectives. *Philosophical Studies*, 57 (3), 261-279. [10.1007/BF00372697](https://doi.org/10.1007/BF00372697)
- Lepore, E. & Smith, B. (Eds.) (2006). *The Oxford handbook of philosophy of language*. Amsterdam, NL: Elsevier.
- Maibauer, J. (2010). Implicature. In A. Barber & R. Stainton (Eds.) *Concise encyclopedia of philosophy of language and linguistics* (pp. 308-321). Amsterdam, NL: Elsevier.
- McGlynn, A. (2012). Interpretation and knowledge maximization. *Philosophical Studies*, 160 (3), 391-405. [10.1007/s11098-011-9725-2](https://doi.org/10.1007/s11098-011-9725-2)
- Putnam, H. (1975). The meaning of "meaning". In K. Gunderson (Ed.) *Language, mind, and knowledge* (pp. 131-193). Minneapolis, MN: University of Minnesota Press.
- Searle, J. (1978). Literal meaning. *Erkenntnis*, 13 (1), 207-224. [10.1007/BF00160894](https://doi.org/10.1007/BF00160894)
- (1983). *Intentionality*. Cambridge, UK: Cambridge University Press.
- (1992). *The rediscovery of the mind*. Cambridge, MA: MIT Press.
- Smith, D. W. & McIntyre, R. (1982). *Husserl and intentionality*. Dordrecht, NL: Reidel.
- Stainton, R. (2010). Context principle. In A. Barber & R. Stainton (Eds.) *Concise encyclopedia of philosophy of language and linguistics* (pp. 88-94). Amsterdam, NL: Elsevier.
- Williamson, T. (2000). *Knowledge and its limits*. Oxford, UK: Oxford University Press.
- (2007). *The philosophy of philosophy*. Oxford, UK: Blackwell.

Grasping Meaning

A Commentary on Christian Beyer

Anita Pacholik-Żuromska

Christian Beyer, referring to a combination of Husserl's and Searle's theses, proposes an account of meaning that is context-dependent and that expresses not only propositional content but also the intentional state of the speaker. However, he tries to weaken Searle's Background Hypothesis, which should be restricted only to the speaker. Thus he excludes from the relation of intentional directedness the third element (called either the hearer, interpreter, or consumer). I will argue that if avoiding radical contextualism is right, it cannot be implemented at the cost of the Background Hypothesis and the triadic relation of intentionality.

Keywords

Background hypothesis | Content | Contextualism | Enactive cognition | Enactivism | Extension | Externalism | First-person perspective | Indexical | Intension | Intentionality | Judgment | Knowledge | Knowledge-how | Meaning | Meaning-function | Possession condition of concepts | Propositional attitudes | Self-identification | Sense | Twin earth thought experiment

Commentator

Anita Pacholik-Żuromska

anitapacholik@gmail.com

Uniwersytet Mikołaja Kopernika
Toruń, Poland

Target Author

Christian Beyer

christian.beyer@phil.uni-goettingen.de

Georg-August-Universität
Göttingen, Germany

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

Since the linguistic turn the main problems considered by philosophers of mind and language are the questions of how words connect with the world, what relations exist between words and objects, what makes utterances true or false, and how we can extrapolate propositional content from internal mental states on external reality. These are particular questions that stem from the general issue of meaning. Our target article is concerned with the question of grasping the meaning and intention that stands behind expressions in the process of producing and interpreting assertive utterances. Its author argues for the thesis that meaning is context-de-

pendent, but in order to properly grasp the meaning of utterances one does not need to have knowledge-how, characterized by John Searle in the form of so-called "Background". Instead, the author proposes a neo-Husserlian conception that allows the reading of intentions standing behind assertions, without reference to factors coming from external context—although this is not an internalistic standpoint. However, taking this position he excludes from the relation of intentionality its third element, namely the hearer (interpreter), depriving him of some kind of responsibility for knowledge about factors determining the truthfulness of asser-

tions. He believes that for the hearer to understand literal meaning knowledge-how is not necessary.

This commentary presents four objections against Beyer's arguments about understanding the meaning of sentences and one separate criticism of his approach to the problem of intentionality. At the beginning I shall reconstruct the thesis and arguments of the author. In the following sections the theses of Beyer will be considered in the light of the general question: what does it mean for meaning to be context-dependent? Here the issue of the differences between contextual and literal meaning will be discussed with reference to Searle's Background Hypothesis. The line of the argumentation will rest on four objections to Beyer's claim about the restriction of the Hypothesis, and will focus on: (1) the problem of indexicals; (2) the distinction between literal and contextual meaning; (3) semantic and social externalism; and (4) understanding as epistemic triangle. The last part of the commentary will be concerned with intentionality considered as a triadic relation strongly connected with the model of understanding. This assumption should lead to an answer to the question of why we cannot reduce the requirements of the Background Hypothesis to producers only.

Even at this early stage, according to Beyer's account, we might ask whether, if the interpreter of the article in question was to misunderstand the article, who has made the mistake—the speaker (producer, author) or the hearer (consumer, reader)? This is another open question that shall accompany this commentary.

2 *Précis of Meaning, Context, and Background*

Arguing for a version of meaning that is context dependent, yet still accessible to every competent language user, Beyer combines two standpoints toward the relation between meaning and intentionality in the work of Edmund Husserl and John Searle. Linking the theses of both philosophers, he assumes that:

1. The meaning of assertive utterances is context dependent.
2. Assertive utterances express not only propositional content but also an intentional state.
3. Searle's Background Hypothesis about the requirement of non-intentional background on the part of the speaker and hearer for recognizing intentional states expressed by assertive utterances as well as for grasping the respective meaning of the utterances could be relevant to an understanding of the context-dependency of assertive utterances, but only in a restricted form.

Beyer's main thesis can be summarized as follows:

The speaker uttering a sentence intentionally presents herself as performing or undergoing an act, but if the hearer does not recognize that intention, she does not thereby fail to grasp the literal truth-conditional meaning of an utterance. Hence, only the group of speakers (utterance producers) must meet the requirements of Searle's Background Hypothesis.

In other words, according to Beyer, context dependence does not prevent competent language users who lack the correct background from grasping the literal truth-conditional meaning of an utterance.

Beyer gives brilliant examples, which justify this main claim. The first group contains indexicals like "I", "here", and "now", which share the same general meaning function—which I generally prefer to call "sense" or "concept"—but which have different respective meanings, that is, a different extension. Take an example, in which Subject 1 asserts: "I have blood type A", and Subject 2 also asserts: "I have blood type A". Both utterances have the same general meaning-function, but express different truth-conditional contents—or propositions. Using an alternative philosophical terminology, they have the same intension but different extension, which results in the famous conclusion that intension does not determine extension (Putnam 1975). However, according to Hilary Putnam's Twin Earth Thought Experiment, even natural kinds "have an indexical unnoticed component" (1975, p. 152). This forces the con-

clusion that every sentence is somehow context dependent, including those containing concepts of natural kinds.

To the second group of examples belong sentences without established uses, such as have been proposed by Searle: “Bill opened the mountain”; “Sally opened the grass”; “Sam opened the sun”. As Searle claims, in the case of such sentences we have no clear idea what they mean, or else we fail to find a proper way of understanding the sentences because we lack the necessary background capacities and social practices.

We know how to open doors, books, eyes, wounds and walls; and the differences in the Network and in the Background of practices produce different understandings of the same verb. Furthermore, we simply have no common practices of opening mountains, grass, or suns. It would be easy to invent a Background, i.e., to imagine a practice, that would give a clear sense to the idea of opening mountains, grass, and suns, but we have no such common Background at present. (Searle 1983, p. 147)

However, Beyer claims that even if we do not have the background we can still grasp the literal meaning of such sentences. We lack knowledge about **verification**—here Beyer agrees with Emma Borg (2004)—i.e., **knowledge-how**, but we can still understand the sentence.

Another example given by Beyer concerns situations where the speaker utters a sentence that the hearer repeats, while referring to another object than that referred to by the speaker. In other words, the hearer mistakenly takes for entitlement¹ an uttered claim about an object, which he thinks is the right referent—for example, when saying “This is red”, the sentence refers to a ball in a box, which the hearer does not know about because he has seen only a

red apple being put into the box. Beyer claims that, according to the principle of knowledge maximization formulated by Timothy Williamson, the speaker should be regarded still as possessing some knowledge about the apple, even if he has a false belief about that object, because even a false judgment in certain circumstances can count as knowledgeable. However, Beyer proposes a modification of this principle, which should, according to him, be “supplemented by a more traditional theory of justification, drawing upon notions of observation, memory and testimony” (Beyer this collection). From the examples given above Beyer infers that contextualism is the right account for this phenomenon, but only in a form that allows minimal semantic knowledge concerning the literal meaning, which can be possessed even in the absence of Background.

3 What does it mean for meaning to be context-dependent?

Epistemic or semantic contextualism has been created as an answer to a sceptical challenge against knowledge in the sense of *episteme*—defined as justified, true belief. It is claimed in this conception that the satisfaction conditions for “x knows that p”—i.e., the truth-conditions of sentences—on whose basis we ascribe knowledge to a subject, depend on the context in which they are uttered, i.e., on epistemic standards obtaining in these contexts (cf. Palczewski 2013, p. 197). “Contextualists speak of the semantic value of knowledge ascriptions as somehow shifting with context [...] The parameter that shifts with the context may be the threshold of justification, the standard of epistemic position, the set of epistemic alternatives” (Preyer & Peter 2005, p. 3).

In contrast to contextualist meaning, for literal truth-conditional meaning we have to look to semantic content. As Searle claims, it is a meaning with “zero context”, determined by the meaning of its semantic components and syntactic rules of composition. However, “for a large class of sentences there is no such thing as the zero or null context for the interpretation of sentences, and that as far as

¹ The *terminus technicus* “entitlement” plays an important role in the philosophy of Robert Brandom, built on the inferential role of the semantic. In Brandom’s account, understanding relies on the participation of subjects in a language game of giving and asking for reasons, where entitlement can be defined as “giving a reason” for repeating the judgment as being true about the object it concerns. I thank Daniel Żuromski for pointing this out.

our semantic competence is concerned we understand the meaning of such sentences only against a set of background assumptions about the contexts in which the sentence could be appropriately uttered” (Searle 1978, p. 207).

The distinction between literal and contextual meaning is clear for Beyer. Literal meaning is not usage. It is a subject of the semantics but not pragmatics. One can grasp the literal meaning of the sentence “The snow is white”, adding to it that this sentence is true only if the snow is white. But when a speaker utters the sentence “The snow is white”, the hearer needs not only to understand the literal meaning, because otherwise he could simply ask “So what?” The hearer needs also to interpret the statement, inferring what kind of linguistic function this sentence fulfils. The hearer needs to understand why (or for what purpose) this statement has been uttered by the speaker. In other words, to grasp the proper meaning he needs to establish what pragmatic and epistemic consequences it has. Thus, the pragmatic consequence is investigated by checking what else the utterance communicates, and what the sentence pragmatically implies. But meaning as usage is not only a matter of implicatures or presuppositions. The epistemic consequences concern the setting of conditions in which the sentence can be truly uttered—that is, the background. To understand an utterance expressing some kind of intentional state, both speaker and hearer have to dispose the background, i.e., knowledge-how. In fact, this is a passive form of knowledge, which depends on physical and social determinants, on which a subject has a little influence and in which she is deeply rooted. Such utterances are evidence of propositional attitudes with certain representational content. In other words, a subject uttering a sentence also expresses (using the terms of folk psychology) his attitude toward its content. However, according to Searle, propositional attitudes are not intentional states, understood as a relation of being directed (or of taking an attitude, i.e., belief) toward a judgment in a logical sense, expressed in the form of a sentence (utterance). “There is indeed a relation ascribed when one ascribes an Intentional to a person, but is not a relation between

a person and a proposition, rather it is a relation of representation between the Intentional state and the thing represented by it. In other words, proposition is rather a content of a statement than its object” (Searle 1983, p. 19).

Searle’s standpoint does not convince Beyer, who claims that to express or correctly ascribe a meaning intention and, consequently, to grasp the literal truth (the conditional meaning of an assertive utterance) one does not need to meet the requirements of Searle’s Background Hypothesis—according to which a subject needs to dispose a set of nonrepresentational capacities to correctly interpret the meaning of utterances. These requirements must be fulfilled only by sentences-producers, who can be regarded as “experts”—however, not necessarily in a scientific sense.

4 Meaning and intentionality

As Beyer claims, if the hearer does not recognize an intention accompanying an utterance, she does not fail to grasp the literal truth-conditional meaning of an utterance. Arguing for this thesis, Beyer gives examples of sentences that do not have an established use or that share the same general meaning function but have different respective meanings. But here are some objections:

The first question concerns indexicals: could we really grasp the literal meaning of the indexical “I” if we could not dispose a background of self-identification? In other words, what would be the distinctive features of context that allow the right ascription of beliefs, if subjects A and B utter the same content, and in the same context? It might be, for example, a capacity to identify themselves as subjects of a certain state, which is a capacity belonging to the unintentional background. If we do not dispose a concept of an individual subject, but only of collectivity, self-identification would be disturbed. In that case, could we still grasp the literal meaning of a sentence like “I do x”? Such self-identification depends on many factors—physical, like a completely unintentional sense of proprioception or homeostasis, and social, based on norms and rules. The case of physical

factors determining the ability to self-identify shows that the Background Hypothesis cannot be reformulated such that the Background must contain intentional elements. As [Searle](#) writes:

On the conception I am presenting, the Background is rather the set of practices, skills, habits, and stances that enable Intentional contents to work in the various ways that they do, and it is in that sense that the Background functions causally by providing a set of enabling conditions for the operation of Intentional states. (1983, p. 158)

Intentional elements would not help our grasping of the meaning if they referred to subjective intentions, which, as Beyer admits, are fully accessible only from first-person perspective. Beyer also doubts whether it is possible to make a comprehensive list of assumptions about a hidden object. But Searle's Background Hypothesis was created precisely to avoid such a regress.

The second question concerns the distinction between literal and contextual meaning. Namely we can raise the doubt: if a hearer does not grasp the contextual meaning, i.e., the truth-conditional meaning, then might she only grasp the sense of the utterance, and not its meaning? If we change terminology, and call general meaning function "sense" or "concept", then we could use Frege's theory of sense and meaning (intension and extension) and say that a subject who grasps only conventional linguistic meaning but not respective meaning grasps *de facto* not the meaning of a sentence but its sense. According to Frege's theory of intension/extension of a sentence, one cannot know a sentence's meaning if one does not know its truth-conditions, because the meaning of a sentence is its truth-value ([Frege 1948](#)). Further, if we turned to were Frege as interpreted by Michael Dummett, we could say that a subject who does not know the truth-conditions of some sentence does not understand this sentence, because, according to [Dummett](#), a theory of meaning should be a theory of understanding ([1993](#)).

The third objection can be formulated as follows: if, according to Beyer, only a producer carries the burden of the requirements of the Background Hypothesis, and if she was a false expert, is there a method (also accessible to a hearer who does not have to know the background) for the identification of false experts by a non-expert? This is a version of Putnam's externalism, which says that external factors, which determine the content of our beliefs could be experts, who for example tell us how to properly use the names "elm" and "beech" (cf. [Putnam 1975](#), p. 145). But what if these experts just pretend to be professionals, or simply have a gap in their education?

If only producers should carry the burden of the requirements of the Background Hypothesis, consumers would have limited access to methods enabling the identification of the satisfaction conditions of an uttered sentence. Hence consumers, grasping only literal meaning, would have to believe everything they heard. As was said, intentionality should not be regarded as a feature of an individual mind. Intentionality is a relation between minds and the world. It is a social phenomenon, developed and practiced through interactions with other minds (cf. [Tomassello & Rakoczy 2003](#)). Hence there must be a theory that can explain how both speaker and hearer have a potentially equal chance of understanding a sentence (of grasping its truth-conditional content). Such a model of understanding has been proposed by Christopher Peacocke. Peacocke claims that the thinker can only judge the content that she recognizes (cf. [Peacocke 1992](#), p. 51). Recognition is possible only if the person knows the truth-conditions of the grasped content. According to Peacocke, the basic concepts are individuated by the fact that, in certain circumstances, our beliefs containing these concepts will be true. These beliefs constitute the knowledge of the subject. Peacocke builds his theory on the assumption that components of the propositional content are concepts individuated by their possession conditions, which fix the semantic value of concepts.

The determination theory for a given concept (together with the world in empir-

ical cases) assigns semantic values in such a way that the belief-forming practices mentioned in the concept's possession condition are correct. That is, in the case of belief formation, the practices result in true beliefs, and in the case of principles of inference, they result in truth-preserving inferences, when semantic values are assigned in accordance with the determination theory. (Peacocke 1992, p. 19)

In fact, in such an account, Peacocke's theory of knowledge is a theory of social solidarity, where knowledge is not a privilege and subjects are considered not as monads or individual minds but as creating a new interpersonal subjectivity—i.e., a social sphere. On the basis of Peacocke's model of gaining knowledge, which contains the triadic relation: concepts, the possession condition of concepts, (conditions in which the use of concept is valid), and through semantic value (fixed on the basis of determination theory), this solidarity is possible, because according to this model everyone can verify or falsify judgments of others. I support this account. The so-called "theory of social solidarity" assumes that both speaker and hearer must share the Background in order to have an access to conditions of justification of utterances.

From the third objection follows the next question: if only a producer needs to dispose a background, then what would be an indicator of the proper usage of a sentence? How could a consumer conclude that a producer understands the uttered sentence (that is, is a competent language user)?

As I have suggested, the consumer also has to utilise certain methods to conclude whether the producer understands the uttered sentence. This tool of verification should be the world, as in Donald Davidson's model of epistemic triangulation. In Davidson's theory, meaning is dispositional. He claims that asymmetry, which happens between a speaker and interpreter's knowledge about a word's meaning, is the same kind of asymmetry between the first- second-person perspectives. This means that knowledge about meaning has to be inferential—hence it is to be identified by an interpreter on the basis of

the speaker's behaviour. To understand the behaviour of an agent, the interpreter has to have a hypothesis about her intention, and then check this hypothesis with respect to the external conditions of the world. In this way, he can verify or falsify his interpretation. If it is wrong, then he must change it and form another hypothesis. Interpretation should be undertaken according to a principle of charity, which means that if the hypothesis fails, then it is the probably the interpreter who is wrong and not the sender—here is the place for experts—the interpreter has to assume that the sender acts rationally, but he has tools to prove it (Davidson 1980).

But in the context of the Background Hypothesis we do not even need to refer to Davidson's theory to show the necessity of an external validation indicator. Searle's original account is good enough:

If my beliefs turn out to be wrong, it is my beliefs and not the world which is at fault, as is shown by the fact that I can correct the situation simply by changing my beliefs. It is the responsibility of the belief, so to speak, to match the world, and where the match fails I repair the situation by changing the belief. But if I fail to carry out my intentions or if my desires are unfulfilled I cannot in that way correct the situation by simply changing the intention or desire. In these cases it is, so to speak, the fault of the world if it fails to match the intention or the desire, and I cannot fix things up by saying it was a mistaken intention or desire in a way that I can fix things up by saying it was a mistaken belief. Beliefs like statements can be true or false, and we might say they have the 'mind-to-world' direction of fit. Desires and intentions, on the other hand, cannot be true or false, but can be complied with, fulfilled, or carried out, and we might say that they have the 'world-to-mind' direction of fit. (Searle 1983, p. 8)

As I have emphasized, since Background and Intentionality are strongly connected it is im-

possible to weaken the Background or add intentional elements to it, because then the mechanism of intentional directedness preserving the external and relational character of propositional attitudes will fall. Nevertheless, Beyer rightly begins his considerations with a comparison of the conception of intentionality from Husserl and Searle. What they have common is the antipsychological thesis that intentionality can be expressed in language. Their idea was to separate intentionality from psychological explanations, which is possible when we consider propositional attitudes as reported in sentences containing the I-clause and the that-clause, thus expressing a relation between an attitude and a judgement in a logical sense. In general, anti-psychologists claim that intentionality is a binary relation between mental acts and the world: the contents of mental acts refer to objects, which exist outside of these acts, while the relation of intentionality is represented in sentences. The relational approach to intentionality affects how we think of mental functions and products, such as judging, believing, doubting, and so on, which are themselves relational.

As Beyer underlines, the problem of meaning intention (termed thus by Husserl) concerns the partly subjective nature of experienced content—a factor that creates the content of the proposition associated with the modality of the state and allows the subject to grasp the content of the experienced state. He refers to Franz Brentano, according to whom every conscious mental act is intentional. In other words, consciousness is intentional because it is always a consciousness of something. Consciousness cannot exist without an intentional act of directedness toward itself. This means that characteristic of mental phenomena is their intentionality or the “mental inexistence of an object [and that] every mental phenomenon includes something as object within itself” (Brentano 1973). So, for example, if I hear a sound, I also grasp the phenomenon of hearing. On the other hand, the content of a mental state is characterized as that which can be expressed in an objectively-verifiable judgment, due to the specific nature of the content which allows the subject to move from first-order beliefs to second-order beliefs

that arise when she ascribes to herself a propositional attitude. This switch can be seen as a change in the form of language: from object language referring to the external world to—and here are two possibilities—either metalanguage, in which the subject reports that she has a belief about having a belief, or to subjective language, in which the subject reports having an attitude with a certain content. In the case of metalanguage, this has to do with issues of semantic externalism, like inheriting truthfulness by second-order beliefs.

Meaning exists only where there is a distinction between intentional content and the form of its externalization, and to ask for the meaning is to ask for an Intentional content that goes with the form of externalization. (Searle 1983, p. 28)

Since propositional attitudes are mental states with propositional content, to interpret them correctly one has to dispose a background of physical and social determinants of the content of the sentences expressing the propositional attitude. This is why a proper theory of intentional directedness should treat both speaker and hearer equally. Speaker and hearer cannot be separated. They are so strongly connected that they should be considered holistically as a single intentional structure or one structure of intentional directedness. Only then arises social intersubjectivity, which does not consist only of individual minds but also of interactions between minds and world as in Davidson's model of triangulation. This relation works for both sides. And the constitution of an individual self is an effect of switching between individual and social minds and between the beliefs of these two kinds: social and individual. It happens for example when an individual mind joins a group and meets regularities different to her own (cf. Tomasello et al. 2005). This means that sometimes, for some reason, it is useful for her to change her beliefs or even her belief-system. She must do this on the basis of her own inferences, so she has to have a reason to do it. Done in any other way she would have problems with understanding this new beliefs.

Hence the triadic model of intentional reference contains a structure that simulates relations of understanding between sender, interpreter, and the world. Subjects never live a solitary life, as is claimed by Husserl. That is why the case of intentionality does not concern a solitary mind. This standpoint gives a straightforward route to contemporary theories of enactive cognition, where a subject is embedded in an environment and, to gain knowledge, has to act and interact with the world of objects and other subjects. This point of view, however, leaves little room for epistemological internalism and thus for the Cartesian mind. Followers of theories of enactivism would say that the content of a subject's mental states is deeply rooted in the body's interactions with the environment—because the whole of cognition is.

That is why when one investigates the content of mental states one needs to refer to both the situation and the situated cognizer taken together as a single, unified system (Wilson 2002). Such enactive theories will be a kind of new version of active externalism, which assumes that “the content-fixing properties in the environment are active properties within a sensorimotor loop realized in the very present” (Metzinger 2004, p. 115). This standpoint, however controversial in the light of classic externalism, has much in common with proponents of this view. So, for example, diachronic externalism holds that the causal story, namely all facts in the past that have had an influence on the thinker, together with an environment, are important determinants of the content of a thinker's propositional attitudes. In contrast to this, synchronic externalism holds also that the content of propositional attitudes is determined by the current environment of the thinker and his disposition to respond to it. On the other hand, social externalism holds that the content of thoughts is determined in part by the social environment of a thinker, and especially by how others in our linguistic communities use words. These “others” could be experts, who establish the scientific names of objects, such as, for example, trees. This version of social externalism could prove fruitful when we consider Searle's Background Theory, but it creates trouble for

Beyer. As I have argued above, in the third objection, externalism is the right approach but it is possible only under the condition of the equal treatment of both participants of the communication process, namely the speaker and hearer, and only when they have access to the background.

5 Conclusion

To conclude, the idea of neo-Husserlian approach to meaning combined with Searle's Background Hypothesis seems to be promising. However, there are several questions that need to be answered. The main problem seems to be the postulated restriction of the hypothesis by adding intentional elements and an abolition of its requirements for a hearer. It would be then a new hypothesis, and rather more Husserlian than Searlian. These requirements may impair the triadic relation of intentional reference, which has to remain triadic if we do not want to come back to idea of a Cartesian mind.

I have raised four objections to Beyer's claim about the restriction of the Hypothesis, concerning the problem of indexicals, the distinction between literal and contextual meaning, semantic and social externalism, and understanding as an epistemic triangle. In the first objection about the use of indexical “I” we have asked whether we could really grasp the literal meaning of the indexical “I” if we didn't have a background of self-identification. I have argued that in the proper use of the pronoun “I” we need a special, non-intentional background. The second objection concerned the problem of whether a hearer, who does not grasp the contextual meaning, grasps only the sense of utterance but not its literal meaning. Answering this question, I claimed that in some approaches—such as, for example, the Dummettian version of Frege's sense and meaning—a subject who does not know the truth-conditions of some sentence does not understand the sentence. The third and fourth objection concerned the restricted role of the hearer in the act of communication. I raised a doubt about whether it is possible to identify false experts and to recognize incompetent language users if the hearer (interpreter)

lacks a non-intentional background. I claimed that to do this, the relation of intentionality must contain three elements: speaker, hearer, and world, where both hearer and speaker have equal access to the background. The relation of intentionality has been considered to be strongly connected with the model of understanding, where speaker and hearer make one unified structure of intentional directness. In such an account, the requirements of the Background Hypothesis cannot be restricted solely to producers, as Beyer would have it.

References

- Beyer, C. (2015). Meaning, context, and background. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Borg, E. (2004). *Minimal semantics*. Oxford, UK: Oxford University Press.
- Brentano, F. (1973). *Psychology from an empirical standpoint*. London, UK: Routledge & Kegan Paul.
- Davidson, D. (1980). *Essays on actions and events*. Oxford, UK: Oxford University Press.
- Dummett, M. (1993). *The seas of language*. Oxford, UK: Clarendon Press.
- Frege, G. (1948). Sense and reference. *The Philosophical Review*, 57 (3), 209-230.
- Metzinger, T. (2004). *Being no one. The self-model theory of subjectivity*. Cambridge, MA: MIT Press.
- Palczewski, R. (2013). Sceptycyzm a kontekstualizm. In R. Ziemińska (Ed.) *Przewodnik po epistemologii*. Kraków, PL: Wydawnictwo WAM.
- Peacocke, C. (1992). *A study of concepts*. Cambridge, MA: MIT Press.
- Preyer, G. & Peter, G. (2005). *Contextualism in philosophy: Knowledge, meaning and truth*. Oxford, UK: Clarendon Press.
- Putnam, H. (1975). The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science*, 7, 131-193.
- Searle, J. (1978). Literal meaning. *Erkenntnis*, 13 (1), 207-224. [10.1007/BF00160894](https://doi.org/10.1007/BF00160894)
- (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge, UK: Cambridge University Press.
- Tomasello, M., Carpenter, M., Call, J., Behne, T. & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28 (5), 675-691. [10.1017/S0140525X05000129](https://doi.org/10.1017/S0140525X05000129)
- Tomasello, M. & Rakoczy, H. (2003). What makes human cognition unique? From individual to shared to collective intentionality. *Mind & Language*, 18 (2), 121-147. [10.1111/1468-0017.00217](https://doi.org/10.1111/1468-0017.00217)
- Wilson, M. (2002). The six views of embodied cognition. *Psychonomic Bulletin & Review*, 9 (4), 625-636. [10.3758/BF03196322](https://doi.org/10.3758/BF03196322)

Self-identification, Intersubjectivity, and the Background of Intentionality

A Reply to Anita Pacholik-Żuromska

Christian Beyer

Two suggestions by Pacholik-Żuromska, concerning the background of “I”-references and the intersubjective dimension of intentionality, respectively, are taken up and related to Husserl’s theory of intentionality. Moreover, a number of misunderstandings of my view are corrected, Searle’s “regress argument” for the Background Hypothesis is criticized, and a distinction between two functions of the background of intentionality is drawn in order to clarify my view.

Keywords

Background hypothesis | Consciousness | Enactivism | Environment | Intentionality | Interactionism | Interpretation | Intersubjectivity | Meaning | Self-identification | Solicitation

Author

[Christian Beyer](#)

christian.beyer@phil.uni-goettingen.de

Georg-August-Universität
Göttingen, Germany

Commentator

[Anita Pacholik-Żuromska](#)

anitapacholik@gmail.com

Uniwersytet Mikołaja Kopernika
Toruń, Poland

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

Pacholik-Żuromska takes issue with both my proposal to tone down Searle’s Background Hypothesis in terms of the distinction between producers and mere consumers and my claim that part of the background of intentionality is itself intentional (albeit in a derived sense). In her introduction she kindly raises the question “who has made the mistake—the speaker (producer, author) or the hearer (consumer, reader),” provided that “the interpreter of the

article [...] was to misunderstand the article” ([Pacholik-Żuromska this collection](#), pp. 2–6). I answer that in case of doubt it is the author of the target article, of course, who has made the mistake. As the formulation of her question shows, Pacholik-Żuromska has indeed misunderstood a central distinction, i.e., that between producer and mere consumer. However, before I take the opportunity to correct this and other misunderstandings, I would like to comment on

two ideas and suggestions by Pacholik-Žuromska that I find interesting and well worth pursuing further.

2 The background of self-identification

The first suggestion concerns our ability to “grasp the literal meaning of the indexical ‘I’” (Pacholik-Žuromska [this collection](#), p. 4) Pacholik-Žuromska contends that this may require, on the part of both speaker and hearer, “a capacity to identify themselves as subjects of a certain state, which is a capacity belonging to the unintentional background.” (Pacholik-Žuromska [this collection](#), p. 4) I agree that (1) the same sort of capacity is in play with both speaker and hearer, and that (2) this capacity belongs to the background. I reject the claim, though, that this capacity is non-intentional. Let me explain.

Ad (1): In *The Thought* Gottlob Frege contends that only the speaker herself can grasp the proposition expressed by the sentence “I have been wounded,” as used in a soliloquy, and that the hearer therefore has to grasp a different proposition, provided by the utterance context, in order to understand a corresponding sentential utterance, such as the proposition expressed by “She who is speaking to you at this moment has been wounded” (Frege 1956, p. 298). This flies in the face of the Husserlian conception of linguistic communication from which I start out in my article, which requires that the hearer ascribes the right meaning-bestowing act to the speaker in a case of successful communication; which means, in the case at hand, that he ascribes to the speaker what Pacholik-Žuromska aptly calls a self-identification (rather than an act of speaking to the hearer, as Frege has it). In fact, this is precisely the way Husserl himself describes what happens in the case of the correct interpretation of “I”-utterances:

Es ist klar: Wer ‘ich’ sagt, nennt sich nicht nur selbst, sondern er ist sich dieser Selbstnennung auch als solcher bewußt, und dieses Bewußtsein gehört wesentlich mit zum Bedeutungskonstituierenden des Wor-

tes ‘ich’. Das aktuelle Sich-selbst-Meinen fungiert [...] so, daß darin sein Gegenstand als Gegenstand eines Selbstmeinens gemeint [...] ist. [...] Der Hörende versteht es, sofern es ihm Anzeige für dieses ganze Bewußtseinsgebilde ist, also der Redende für ihn als jemand dasteht, der sich selbst, und zwar als ‘ich’ nennt, d.i. sich als Gegenstand seiner als Selbsterfassung erkannten Selbsterfassung nennt.¹ (Husserl 1984, p. 813)

Thus, if the speaker asserts “I have been wounded,” she presents herself as someone who refers to herself *as* referring to herself (or as meaning herself/having herself in mind/thinking of herself), in order to state about herself that she has been wounded; and the hearer understands this assertion if he takes the speaker to refer to herself as referring to herself and to assert about herself that she has been wounded. I regard this metarepresentational view of the meaning-bestowing acts underlying the assertive use of “I”-sentences as quite plausible. After all, if someone claims, say, “I have a broken leg,” then she *eo ipso* knows that she *refers to herself* by “I;” she could instantly add: “I am speaking of *myself*.” (Contrast this to a case in which a speaker unknowingly looks at herself in a mirror and exclaims “She has a broken leg.” See Beyer 2006, pp. 33 ff.) Incidentally, this view fits in well with a dispositionalist higher-order judgment theory of consciousness, which implies that (thanks to an underlying, “pre-reflective” structure of inner time-consciousness) “I”-awareness disposes its subject to judge that she herself is thinking of herself (see Beyer 2006, pp. 33 ff.).² If a mental disposition such as this is actualized (which is not re-

1 The English translation is as follows: “Clearly, if someone says ‘I’, he does not only refer to himself, but he is also aware of this referring to himself as such, and this awareness builds an essential part of what constitutes the meaning of the word ‘I’. The current act of meaning oneself is functioning [...] in such a way that in the course of it its [intentional] object is [...] being meant *as* the object of an act of meaning oneself. [...] The hearer understands it, if he takes it as an indication for the whole structure of consciousness just described, that is to say, if the speaker is regarded by him as someone who refers to himself precisely as ‘I’, i.e., as someone who refers to himself as the object of his recognition of himself recognized as a recognition of himself” (My translation.)

quired for self-identification), then the resulting (“reflective”) “I”-judgment is based upon, and epistemically motivated by, a (“pre-predicative”) act of referring to oneself *as* referring to oneself.

Ad (2): In order for the hearer to ascribe such a meaning-bestowing act of self-identification to the speaker, he does not, of course, have to actualize his own capacity for self-identification, in the sense of actually thinking of himself as thinking of himself. But in the absence of this *capacity* he would be unable to ascribe such an act to the speaker, at least if we follow Husserl and Edith Stein and conceive of third-person act ascriptions as based on empathy, where the ascriber mentally simulates the cognitive situation of the target person (see [Beyer 2006](#), ch. 3). So I agree with Pacholik-Żuromska that an element of the background (notably, the capacity for self-identification) is required for the ability to understand “I”-utterances. However, I deny that this capacity is completely non-intentional. It is precisely a mental disposition that is actualized (if it gets actualized) in *intentional* consciousness, namely in pre-predicative acts of referring to oneself *as* oneself—acts which may, but need not, give rise to corresponding “self-reflective” higher-order judgments.

3 The intersubjective dimension of intentionality

Another interesting suggestion made by Pacholik-Żuromska concerns the relationship between intentionality and intersubjectivity. She claims that “[i]ntentionality is a relation between minds and the world” ([Pacholik-Żuromska this collection](#), p. 5), thus subscribing to an externalist conception of intentionality (which I share), and goes on to characterize it as “a social phenomenon, developed and practiced through interactions with other minds” ([Pacholik-Żuromska this collection](#), p. 5). [Pacholik-Żuromska](#) refers to Tomasello, Rakoczy, and Davidson in this connection, but (her ascription to Husserl of the thesis that

subjects can “live a solitary life” notwithstanding, which I regard as a misreading; [this collection](#), p. 8) she could also have referred to Husserl here. In the second volume of his *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy (Ideas II)*, Husserl presents a detailed analysis of the intersubjective, reciprocal constitution of intentional objects that belong to a “communicative environment” and are thus immediately perceivable as valuable heating material, for example:

Kohle z.B. sehe ich als Heizmaterial; ich erkenne es und erkenne es als dienlich und dienend zum Heizen, als dazu geeignet und dazu bestimmt Wärme zu erzeugen. [...] Ich kann [den brennbaren Gegenstand] als Brennmaterial benutzen, [...] er ist mir wert mit Beziehung darauf, daß ich Erwärmung eines Raumes und dadurch angenehme Wärmeempfindungen für mich und andere erzeugen kann. [...] [A]uch andere fassen ihn so auf, und er erhält einen intersubjektiven Nutzwert, ist im sozialen Verbande geschätzt und schätzenswert als so Dienliches, als den Menschen Nützlichendes usw. So wird er nun unmittelbar ‘angesehen’ [...].³ ([Husserl 1952](#), p. 187)

Notice that near the end of this passage from § 50 of *Ideas II* Husserl observes how intersubjective agreement in the form of reciprocally shared emotional valuations, and accordingly motivated evaluations (evaluative judgments), add a social dimension to the constitution of the environment. In this way, the personal environment of an individual subject acquires the significance of a social environment equipped with

3 The English translation is as follows: “I see coal as heating material; I recognize it and recognize it as useful and as used for heating, as appropriate for and as destined to produce warmth. [...] I can use [a combustible object] as fuel; it has value for me as a possible source of heat. That is, it has value for me with respect to the fact that with it I can produce the heating of a room and thereby pleasant sensations of warmth for myself and others. [...] Others also apprehend it in the same way, and it acquires an intersubjective use-value and in a social context is appreciated and is valuable as serving such and such a purpose, as useful to man, etc. That is how it is first ‘looked upon’ in its immediacy.” ([Husserl 1989](#), pp. 196f.)

2 This may also fit in with the Brentanian conception of consciousness that Pacholik-Żuromska alludes to in section 4.

common objects possessing intersubjectively shared values—in the case at hand: shared use-values, to be perceived immediately (e.g., as a piece of heating material). In the following section, §51, entitled “The person in personal associations,” Husserl generalizes these observations. He claims that the social environment is relative to persons who are able to “communicate” with one another, i.e., to “determine one another” by performing actions with the intention of motivating the other to display “certain personal modes of behavior” on his grasping that very communicative intention (Husserl 1989, p. 202). If an attempted piece of communication such as this, also called a “social act” (Husserl 1989, p. 204), is successful, then certain “relations of mutual understanding (Beziehungen des Einverständnisses)” are formed (Husserl 1989, p. 202):

[A]uf die Rede folgt Antwort, auf die theoretische, wertende, praktische Zumutung, die der Eine dem Anderen macht, folgt die gleichsam antwortende Rückwendung, die Zustimmung (das Einverstanden) oder Ablehnung (das Nicht-einverstanden), ev. ein Gegenvorschlag usw. In diesen Beziehungen des Einverständnisses ist [...] eine einheitliche Beziehung derselben zur gemeinsamen Umwelt hergestellt.⁴ (Husserl 1952, pp. 192-193)

A few lines later, Husserl even claims that relations of mutual understanding help determine the common surrounding *world* for a group of persons; the world as constituted this way is called a “communicative environment.” On his view, the world of experience is partly structured by the outcomes of communicative acts. If it is structured this way, then there will be meaningful environmental “stimuli,” or solicitations (to use a more recent terminology), which motivate a given subject to display personal behaviour that consists in his reacting upon such environmental stimuli;

where the notion of motivation is to be understood as follows:

[W]ie komme ich darauf, was hat mich dazu gebracht? Daß man so fragen kann, charakterisiert alle Motivation überhaupt.⁵ (Husserl 1952, p. 222)

I regard this Husserlian conception of the structures underlying our being-in-the-world as highly plausible. So Pacholik-Żuromska kicks at an open door when she stresses the importance of (what is nowadays called) embedded cognition and dynamic mind-world interaction for an adequate conception of intentionality.⁶

4 Some corrections and clarifications

Finally, some corrections. I begin with two misunderstandings that I find easily comprehensible.

First, my use of the term “producer” may be misleading, as it differs from the ordinary use of the term. Not every producer of an utterance, in the ordinary sense, is a producer in the technical sense that Evans and I associate with the term. To take up the example that Evans gives in the long quotation cited at the beginning of section 5 of my article, if someone uses the name “Livingston” today to refer to (say) an 18th century politician, then she will be a mere consumer of that name, because she could not “have been introduced to the [name-using] practice via [her] acquaintance with” Livingston, to put it in Evans’ terms (1982, pp. 376–393). This holds true even if she is the *speaker* of an utterance in which the name “Livingston” is used this way. I do not think that the producer/consumer distinction leads to a problematic

⁵ The English translation is as follows: “How did I hit upon that, what brought me to it? That questions like these can be raised characterizes all motivation in general.” (Husserl 1989, p. 234; in part my translation)

⁶ Pacholik-Żuromska also refers to Davidson’s notion of triangulation in this connection. For a Husserl- and Føllesdal-inspired critique of Davidson’s recourse to causal concepts in this context, see Beyer 2006, pp. 88–99. In the last paragraph of section 4 she draws a distinction between diachronic externalism—a position she ascribes to Davidson—, synchronic and social externalism, claiming that the latter “creates trouble for Beyer” (Pacholik-Żuromska this collection, p. 8). In the light of both the foregoing considerations and her misreading of my view on Searle’s Background Hypothesis (see section 4), I regard this claim as false.

⁴ The English translation is as follows: “[S]peaking elicits response; the theoretical, valuing, or practical appeal, addressed by one to the other, elicits, as it were, a response coming back, assent (agreement) or refusal (disagreement) and perhaps a counterproposal, etc. In these relations of mutual understanding, there is produced [...] a unitary relation of [persons] to a common surrounding world.” (Husserl 1989, pp. 203-204)

two-tier society of linguistic insiders and outsiders, as Pacholik-Żuromska seems to believe. It merely reflects the way proper names and other expressions acquire a particular usage, as a matter of fact. Actually, Pacholik-Żuromska herself draws upon a very similar distinction (but see footnote 16 in the target article, [Beyer this collection](#)) when she talks about experts. Of course, in principle anyone may become an expert regarding the application of any term—although it is difficult, to say the least, to become a producer regarding a proper name whose bearer has passed away a long time ago (see above). If this latter remark is correct (as I think it is), then it is not the case that in general “everyone can verify or falsify judgments of others,” as [Pacholik-Żuromska \(this collection, p. 6\)](#) wants to claim following Peacocke. In some cases (such as the case of proper names whose bearers have passed away) some people—the mere consumers—are in an epistemically underprivileged position.

Since Pacholik-Żuromska mistakenly equates what I call producers with speakers and mere consumers with hearers, she misreads my proposal to tone down Searle’s Background Hypothesis in such a way that only the producers with regard to a given set of sentences need “background know-how regarding the application of those sentences” (as I put it in section 5 in the target article, [Beyer this collection](#)), and her relevant arguments are besides the point—even if they contain interesting ideas (see sections 1 and 2 above). I do not claim that the Background Hypothesis “should be restricted only to the speaker,” as [Pacholik-Żuromska \(this collection, p. 1\)](#) puts it in her abstract. I contend that it should be restricted to the producers.

This brings me to a second misunderstanding that also concerns my view on the Background Hypothesis. In some places (like the last paragraph of section 4 in the target article; [Beyer this collection](#)) I carelessly put my view in such a way that it invites the following interpretation, which Pacholik-Żuromska takes for granted: only the producers need any background know-how. However, this is, again, a misreading, as becomes clear when one looks at

more careful formulations of my view, such as the one quoted in the preceding paragraph or the following formulations from section 5: “Meaning-intentions [...] do not generally require a non-intentional background *relative to which their (truth-conditional) content and satisfaction conditions are determined*,” ([Beyer this collection, p. 15](#); emphasis added) “the applicability of the Background Hypothesis [...] needs to be restricted, *as far as the part of the background (co-)determining truth-conditional content is concerned, to what I have called the producers*.” ([Beyer this collection, p. 17](#); emphasis added) What Pacholik-Żuromska does not notice, and what I should have made clearer, is that I distinguish between two different functions of the background:

- On the one hand, some of its elements (such as personal acquaintance with a name-bearer, or with a practice like opening a can) help to determine a particular truth-condition for a sentence-use and its underlying meaning-bestowing act—here I claim that only the producers need a corresponding background.
- On the other hand, the existence of what Searle calls the Network is an enabling condition for intentional consciousness.

Regarding the latter, I argue near the end of my article that it is misleading to characterize the part of the Network that constitutes “the set of anticipations determining” ([Beyer this collection, p. 16](#)) what Husserl refers to as the “intentional horizon” of a conscious intentional state as completely non-intentional, because they are mental dispositions to form occurrent higher-order beliefs. In order to save Husserl’s notion of intentional yet unconscious horizon anticipations, which I regard as an indispensable contribution to the theory of intentionality, I propose that we (re)formulate Searle’s background conception in such a way that “the background may indeed contain intentional elements, albeit in a derived sense” ([Beyer this collection, p. 17](#)), notably in the sense of mental dispositions to form higher-order beliefs. The only argument I find in Pacholik-Żuromska’s commentary that may at

first sight be taken to speak against the admission of such intentional background-elements is the regress argument she refers to in section 4 of her commentary. She points out that the Background Hypothesis is supposed to avoid a regress of assumptions such as the one I describe in section 3 in the target article (under the heading “Background assumptions”) in order to motivate Searle’s radical contextualism. But, quite apart from the fact that I do not claim that *all* elements of the background are intentional in the relevant sense, I find Searle’s corresponding argument for the Background Hypothesis confused. He claims that “[t]he actual content is insufficient to determine the conditions of satisfaction,” and that “[e]ven if you spell out all the contents of the mind as a set of conscious rules, thoughts, beliefs, etc., you still need a set of Background capacities for their interpretation.” (Searle 1992, pp. 189–190) To repeat a point I make in section 5 of my article, this is an absurd view (cf. Beyer 1997, p. 346). Neither intentional content (“actual content”) nor respective meaning can be interpreted (or “applied”) at all—only (utterances of) linguistic *expressions*, including *formulations* of rules, can be interpreted, and the result of this interpretation will be (the ascription of) a meaning-bestowing act which displays an intentional content that uniquely determines the conditions of satisfaction.

Here are some further corrections and clarifications.

I do not take the case of indexicals like “I,” “here” and “now” to show that “literal truth-conditional meaning” can be grasped even in the absence of “the correct background.” Unlike “sentences without established use” (as Pacholik-Żuromska aptly calls them; [this collection](#), pp. 2–3), these examples have no bearing on the truth of the version of the Background Hypothesis for which I argue. They can be captured by any standard semantics that distinguishes between general meaning-function (character) and respective meaning (content).

I do not give any example in which “the speaker utters a sentence that the hearer re-

peats, while referring to another object” ([Pacholik-Żuromska this collection](#), p. 3) than the one the speaker refers to. In the example about the yellow apple and the red ball in the box, the speaker refers to the apple in order to (wrongly) state that it is red, and the hearer may figure this out by applying a suitably modified version of Williamson’s principle of knowledge maximization (rather than Davidson’s principle of truth maximization). Nor do I claim that any “false judgment in certain circumstances can count as knowledgeable.” ([Pacholik-Żuromska this collection](#), p. 3) Rather, the unmodified version of Williamson’s principle is not applicable in the case at hand.

Furthermore, epistemic contextualism does not only (often) purport to answer sceptical challenges to *justified-true-belief* accounts of knowledge, but also to other accounts such as reliabilist theories of knowledge that make recourse to the notion of the ability to exclude relevant alternatives. I do not distinguish between “literal truth-conditional meaning” and “contextual respective meaning,” as Pacholik-Żuromska claims in section 3. In the case of indexicals, literal meaning is not to be confused with linguistic meaning in the sense of general meaning-function (character). The relevant distinction is that between literal and figurative meaning; unlike “meaning as usage,” figurative (or non-literal) meaning is a case of (what is expressed by) implicature.

It is misleading to assert that “according to Searle, propositional attitudes are not intentional states” ([Pacholik-Żuromska this collection](#), p. 4). It is true, however, that (like Husserl) Searle does not conceive of them “as a relation of being directed [...] towards a judgment in a logical sense” ([Pacholik-Żuromska this collection](#), p. 4), i.e., towards a proposition. Propositional contents are to be distinguished from the satisfaction conditions they determine (which Husserl refers to as states of affairs).

I do not claim that “if the hearer does not recognize an intention accompanying an utterance, she does not fail to grasp the literal truth-conditional meaning.” ([Pacholik-Żuromska this collection](#), p. 4) Grasping that mean-

ing requires, on the part of the hearer, to ascribe the intention to express a meaning-bestowing act to the speaker.

I distinguish (following Borg) between knowing a sentence's truth-condition, on the one hand, and being able to decide whether this truth-condition is satisfied, on the other. Mere consumers *do* know the truth-condition of a sentence when they understand it, but they are unable to decide (in an epistemically responsible way) whether it is met. *Pace Pacholik-Żuromska*, this does not mean that they "have to believe everything they [hear]." ([this collection](#), p. 5) Nothing in the notion of a mere consumer implies that he must regard a given speaker as infallible (and sincere), even if this speaker is in fact a producer; and nothing in the notion of a producer implies that producers are infallible (and always truthful). Nor does this mean that producers grasp truth-conditional content more "fully" than mere consumers (see footnote 16 of my target article; [Beyer this collection](#)).

On my conception of a producer, there can be no producer who is "a false expert." It is possible, though, on my view, to be a producer without being a scientific expert on whatever it is that constitutes the extension, reference, or truth-condition of the relevant expression (again, see footnote 16 of my target article; [Beyer this collection](#)).

Pacholik-Żuromska raises an excellent question when she asks what, on my view, "would be an indicator of the proper usage of a sentence." ([this collection](#), p. 6) However, it does not speak against a particular approach to meaning that this problem arises in its framework. It arises in any framework.

Husserl took over the idea of intentionality from Brentano, but he does not share Brentano's view that consciousness is always intentional. According to Husserl, there is also non-intentional consciousness, such as pain. Without intentionality, there would be no stream of consciousness, and hence no consciousness. But not every single element of the stream of consciousness is itself intentional. As usual, I find myself in agreement with Husserl here.

5 Conclusion

Despite some serious misunderstandings, for which I am prepared to take responsibility at least in part, *Pacholik-Żuromska* presents some promising ideas. In particular, she highlights the significance of the background of self-identification and the intersubjective dimension of intentionality. In addition, her commentary has helped me to see the need to explicitly distinguish between two functions of the background: its reference-determining role on the one hand, and its enabling role in connection with the functioning of the intentional horizon on the other.

References

- Beyer, C. (1997). Husserl's representationalism and the "hypothesis of the background". *Synthese*, 112 (3), 323-352. [10.1023/A:1004992424269](#)
- (2006). *Subjektivität, Intersubjektivität, Personalität*. Berlin, GER: De Gruyter.
- (2015). Meaning, context, and background. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Evans, G. (1982). *The varieties of reference*. Oxford, UK: Clarendon Press.
- Frege, G. (1956). The thought: A logical inquiry. *Mind, New Series*, 65 (259), 289-311. [10.2307/2251513](#)
- Husserl, E. (1952). *Ideen zur einer reinen Phänomenologie und phänomenologischen Philosophie. Zweites Buch: Phänomenologische Untersuchungen zur Konstitution*. Den Haag, NL: Nijhoff.
- (1984). *Logische Untersuchungen 2. Teil*. Dordrecht, NL: Nijhoff.
- (1989). *Ideas pertaining to a pure phenomenology and to a phenomenological philosophy: Second book*. Dordrecht, NL: Kluwer.
- Pacholik-Żuromska*, A. (2015). Grasping meaning—A commentary on Christian Beyer. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Searle, J. (1992). *The rediscovery of the mind*. Cambridge, MA: MIT Press.

The Puzzle of Perceptual Precision

Ned Block

This paper argues for a failure of correspondence between perceptual representation and what it is like to perceive. If what it is like to perceive is grounded in perceptual representation, then, using considerations of veridical representation, we can show that inattentive peripheral perception is less representationally precise than attentive foveal perception. However, there is empirical evidence to the contrary. The conclusion is that perceptual representation cannot ground what it is like to perceive.

Keywords

Acuity | Adaptation | Appearance | Attention | Awareness | Consciousness | Content | Contrast | Endogenous attention | Exogenous attention | Grounding | Indeterminacy | Marisa Carrasco | Perception | Peripheral perception | Precision | Reductionism | Representational content | Representationism | Saliency | Tyler Burge | Unconscious perception | Vagueness | Veridicality | Visual field

Author

Ned Block

ned.block@nyu.edu

New York University

New York, NY, U.S.A.

Commentator

Sascha B. Fink

sfink@ovgu.de

Otto von Guericke Universität

Magdeburg, Germany

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität

Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University

Melbourne, Australia

1 Introduction

Attention increases acuity, allowing the perceiver to see details that would otherwise be missed. In addition, for items that the perceiver does actually see, attention changes their appearance, increasing, for example, the appearances of contrast, (differences between light and dark), speed of a moving object, spatial frequency (a measure of how closely spaced light and dark areas are) and the size of a gap—as in Figure 4. But when attention makes something appear bigger or faster, does it work like a magnifying glass, trading off a gain in information at the cost of making something appear bigger or faster than it is? Or does attentive perception portray the item more as it really is? Or are both percepts veridical—or are both non-veridical? Similar issues arise with regard to inhomogeneities in the visual field. Vision in the

lower visual field is about 65% more sensitive to contrast (and orientation discrimination, texture segmentation, gap size, speed, spatial frequency) than vision equidistant from fixation in the upper visual field. (See Figure 1 for examples of low and high contrast.) In addition, there is a great deal of noise in perceptual systems. Percepts involving the same area of the visual field and the same degree of attention will typically differ in visual response from occasion to occasion. So on different occasions, one can see the same object or event in the same conditions, with the same degree of attention, and from the same vantage point and it will look different in size or speed or contrast because of random factors.

What is the consequence of these facts for the veridicality of perception? One viewpoint



Figure 1: Six levels of contrast. The Wikipedia caption reads “Different levels of contrast - original image top left - less contrast to the left (50%, 75%), more to the right (25%, 50%, 75%)”. I take this to mean that the mid-left photo has 50% less contrast than the upper left, the lower left photo as 75% less contrast than the upper left, etc. These percentages are differences from photoshop, not absolute measures of contrast of the sort to be discussed later in the paper. Percent contrast in the sense to be discussed is the difference between the luminance of the lightest and darkest parts divided by the sum of these luminances. These images come from the Wikipedia entry on contrast. According to Wikipedia, “Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled *GNU Free Documentation License*.”

says that perception is mostly slightly mistaken. We usually see length, speed and contrast non-veridically but the extent of error is small enough not to be problematic. However, this viewpoint cannot be right since it is only in virtue of a history of veridical representation both in our own lives and in the past of our species that our perceptual representations even have representational contents (Burge 2010). Without such a history of veridical representation it is not clear that perceptual representation really makes sense.

An alternative way of thinking about the issue is that perception is sufficiently imprecise in its representational content for all these varying percepts to be veridical. If a person is said to be 5 feet to 6 feet tall on one occasion and 6 feet to 7 feet tall on another, both are veridical if the person is 6 feet tall. One could put this by saying that perceptual representation is “intervallic”. The intervals however would have to be pretty large given the size of these effects—notably the 65% difference between lower and upper visual field just mentioned. And it is hard to square such large differences with the phenomenology of foveal vision. Hold a piece of lined paper in front of you. You seem to see the difference between the white space and the lines fairly precisely. “Irrelevant!,” you may retort, “Those differences in the visual field affect only peripheral perception; attentive foveal perception is much more precise than inattentive peripheral vision.” And this resolution seems to be reflected in our phenomenological judgments: move the piece of lined paper out to 30° away from the line of sight. Doesn’t your visual impression of the contrast between the lines and spaces seem, well, less precise? Surprisingly there is evidence that unattended and peripheral perception of some properties (notably contrast) are about as precisely represented in attentive foveal vision as in inattentive vision and vision in the near periphery (up to a 30° angle from the line of sight). The upshot is that the phenomenology of perception may mislead us with regard to the precision of the representational content of perception.

One might suppose that help will come from bodily action. Goodale & Murphy

presented 5 rectangular blocks to subjects at various positions in the visual field ranging from 5° to 70° away from the line of sight (1997). They compared accuracy of perceptual discrimination of one block from another with accuracy of grip via a device that measured the aperture between thumb and forefinger as subjects reached out to pick up one of the blocks. Grip accuracy is roughly the same at 5° as at 70°. The fine details of action are controlled by a largely distinct system from the system that underlies conscious vision. So what this result dramatically illustrates is that the precision of bodily action is unlikely to cast any light on the precision of perceptual phenomenology.

This is the puzzle of the title. I argue that the disconnect may be real and that perceptual phenomenology may mislead about perceptual representation. Perceptual phenomenology may not be grounded in the representational content of perception. Further, there may be no “phenomenal content”, that is no representational content that emerges from the phenomenology of perception.¹

This is a very long paper so it might be useful to know what parts to focus on. You can see the basic lines of the dialectic from reading sections 1-3. Sections 4-7 concern the experimental data concerning attention and can be skimmed without losing the thread. The argument resumes with 8-10. 11 can be skipped without loss of continuity. 12 covers some of the results that the argument is based on. 13 can be skipped. 14 is the conclusion.

2 Background

This section describes some assumptions and terminology. A simple percept consists of a representation of an environmental property and a singular element that picks out an individual item (Burge 2010). The representational content is the condition of veridicality and is satisfied only if the referent of the singular element has

¹ Direct realists reject representational contents, holding instead that the phenomenology of perception is grounded in what properties one is directly aware of. They face a parallel set of issues with regard to the question of how precise the properties are that one is directly aware of.

the property represented by the property-representation. The precision of a representation—in my terminology—is a matter of the range of values attributed. For example, consider two visual representations of the height of a person, one representing the person as between 5'6" and 6' tall, the other representing the person as between 5'8" and 5'10" tall. The latter has a narrower precision. Precision in the sense used here is not a matter of indeterminacy of borders but rather the size of the range.²

The claim that the precision of a representation is wide is a form of the claim that perception is “intervalic”. There are other measures of perception that are easily confused with precision. One of them is acuity—also known as spatial resolution. Acuity is the ability to resolve elements of stimuli. Common measures in the case of vision are the extent to which the subject can distinguish one dot from two dots, detect a gap between two figures, determine whether a rotating figure is rotating clockwise rather than counter-clockwise, ascertain whether two line segments are co-linear, distinguish a dotted from a solid line or detect which side of a Landolt Square a gap is on. (See Figure 4 for an example of a Landolt Square.)

These and other items of terminology are gathered together in a glossary at the end of the article. Of course other quite different definitions of ‘precision’ and ‘acuity’ are just as legitimate as these. Note in particular that I am not using the notion of precision as the inverse of variance or the notion of precision associated with the predictive coding literature.

Representationists (also known as representationalists and intentionalists) think that what it is like to have a perceptual experience—that is, the phenomenology of perceptual experience—is grounded in the representational content of the perception. (Not that representationists have used the notion of grounding, but I believe that it captures what they have meant.) Representationism is sometimes framed as an identity thesis (e.g., Pautz

2010; Tye 2009): what it is for an experience to have a certain phenomenal character = for it to have a certain representational content. But the identity formulation is inadequate because the phenomenology is supposed to be based in the representational content and not the other way around. Identity is symmetrical. The grounding characterization of representationism avoids this problem since grounding is asymmetrical. To say that perceptual representation grounds perceptual phenomenology is to say that it is in virtue of the representational content of a percept that it has the phenomenology it has. And *in virtue of* is asymmetrical. (See Fine 2012 on the concept of ground and my 2014a for further discussion of grounding in philosophy of mind.)

Representationism is often framed in terms of supervenience: no difference in the phenomenology of perception without a difference in its representational content. But supervenience does not capture a key motivation behind representationism: that the representational content of perception is the source of the phenomenology of perception, that it is in virtue of the representational content of the perception that it has the phenomenology it has. A supervenience formulation would entail that a difference in the precision of phenomenology requires a difference in representational content. However, on a supervenience formulation of representationism it would be a further question whether the phenomenology of perception could increase in precision without a commensurate increase—or even with a decrease—in precision of its representational content. On the grounding characterization, any change or difference in phenomenological precision is dependent on a commensurate change or difference in representational precision.

The grounding formulation of representationism rules out some but not all kinds of multiple realization. Suppose that red₇₈₂ is an example of the most fine-grained color we can experience. And suppose that the representationist theory of the experience as of red₇₈₂ is that this experience is grounded in representation of red₇₈₂. Different experiences as of red₇₈₂ can be realized by different representational states so

² As Tim Williamson noted when some of this material was presented at Oxford, the fuzziness of the borders is vagueness rather than imprecision. Ryan Perkins & Tim Bayne argue against representationism using considerations of vagueness (2013).

long as they all involve the representation of red₇₈₂.

The grounding characterization captures a representation-first view and excludes phenomenology-first doctrines that are often portrayed as representationist. Phenomenology-first views suppose that phenomenology grounds at least some kinds of representational contents (Hill 2009; Kriegel 2011, 2013; Shoemaker 2007). And it also excludes versions of representationism that treat both the phenomenology and representational content of perception as grounded in something else (Chalmers 2006; Siegel 2013). That is a plus for the grounding characterization—distinguishing between fundamentally different points of view. Although I won't talk about this much here, I think the considerations I will be raising will cast doubt on views that phenomenology grounds any kind of representational content.³

The reader may feel that both peripheral and unattended perception are odd and unimportant phenomena that cannot be the test of any theory of perception. However, peripheral unattended perception is ubiquitous. The fovea is the high density center of the retina. If you hold your hand at arm's length, your foveal perception encompasses about double the width of your thumb. Much of perception at any fixation occurs outside that area and a similar point applies to attention. However, even if you think that both peripheral and unattended perception are atypical, you should recognize that atypical cases often are a window into the nature of a phenomenon. The experiment in which a beam of light goes through two slits was crucial in demonstrating a wave aspect of light (Feynman 1988).

³ Some of the philosophers who call themselves “representationists”, for example Michael Tye (2009), have endorsed “object-involving” representational contents. Suppose I am looking at a tomato and having an experience that represents the tomato as being red₇₈₂. You are looking at an exactly similar tomato in identical circumstances and also having an experience that represents it as having red₇₈₂. According to Tye, we are having phenomenally different experiences in virtue of looking at different tomatoes. As Burge has noted in an article on direct realism (2005), there are object-involving phenomenal types (of the sort Tye is talking about), but there are also non-object-involving phenomenal types. Representationism as discussed here is concerned with the latter types. I mentioned in footnote 1 that the same issues about precision arise for direct realism—and the same applies to Tye's view.

3 The inhomogeneous visual field

Although this article is mainly about differences in perception wrought by differences in attention, it will be helpful to start with a discussion of similar issues that arise independently of attention because of the massive inhomogeneities in the visual field. I will discuss the perception of contrast. The visual system is much more sensitive to differences in luminance than to luminance itself and contrast is a matter of luminance differences. (Luminance is a measure of the light reflected from a surface.) Contrast can be defined in a number of different ways, all ways of capturing the average difference in luminance between the light and dark parts of an array. The four patches in Figure 2 have roughly equal apparent contrasts if one is fixating the cross though there is substantial variation among persons in comparative sensitivities in the visual field. But the top patch has a 30% contrast and the bottom patch has a 15% contrast. (To fixate the cross is to point your eyes at it.) Vision in the lower visual field (the South) has about 65% better sensitivity than vision in the upper visual field on average along the “vertical meridian” (the vertical line through the fixation point) for points of equal eccentricity. And sensitivity is better along the horizontal meridian than the vertical meridian, that is East and West have higher sensitivity than points of equal eccentricity in the North and South. This sensitivity advantage is about 63%. Marisa Carrasco suggests that the advantage of the horizontal over vertical meridians probably has to do with the presence of more relevant information on the horizontal meridian (Carrasco et al. 2001). These differences in sensitivity manifest themselves phenomenologically in differences among patches required for equal apparent contrasts. It takes a 30% contrast patch in the North to phenomenologically match a 10% contrast patch in the East at the equal eccentricity depicted in Figure 2. Performance asymmetries along these lines have been observed for gap size, spatial frequency (roughly density of stripes), orientation discrimination, texture segmentation, letter recognition and motion perception. Performance asymmetries of this sort have been shown in comparisons between an on-screen stimulus and a stimulus from the recent

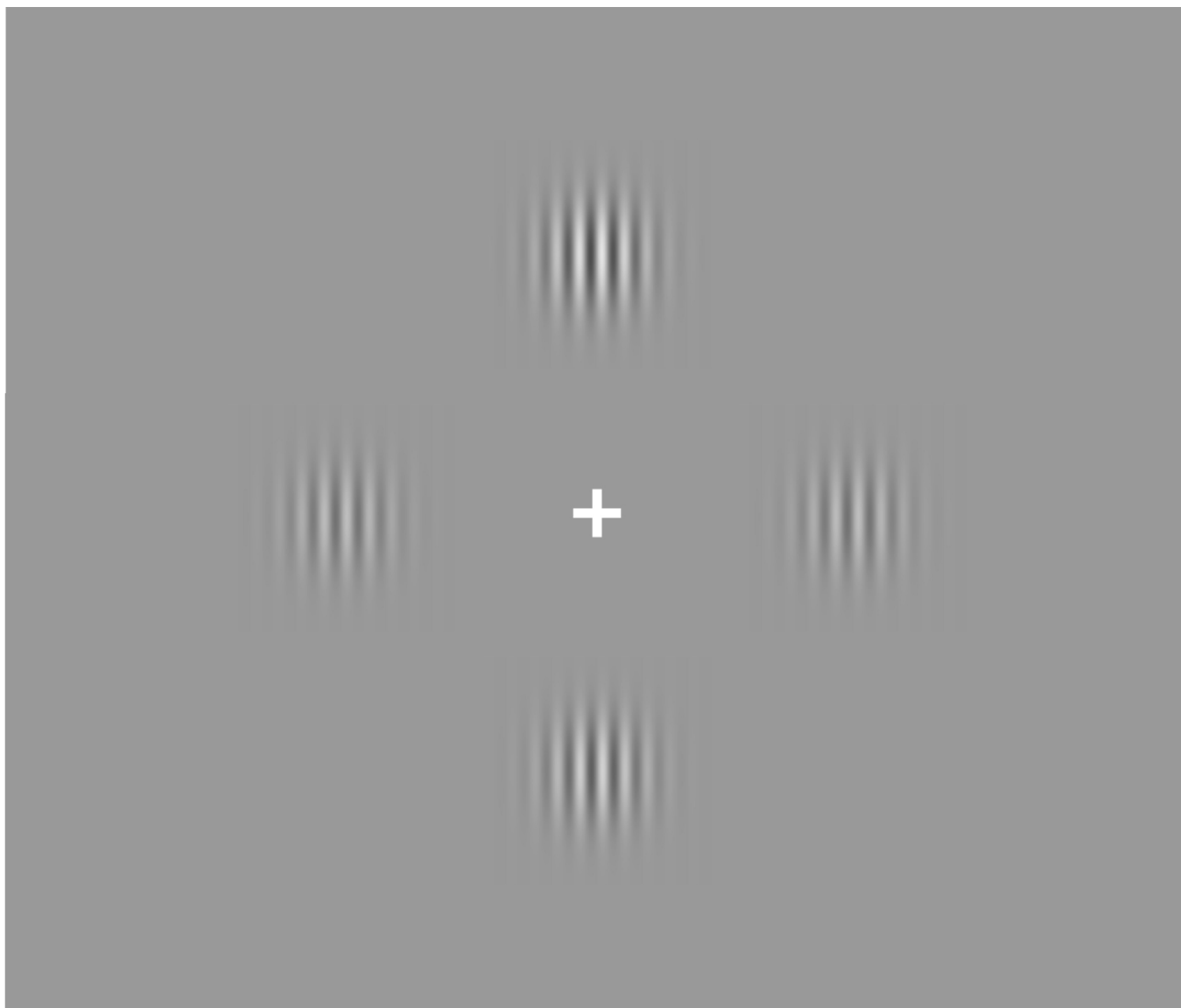


Figure 2: If you fixate (i.e., point your eyes at) the plus sign, these four different patches should look roughly equal in contrast at normal reading distance (roughly 15 inches away). The one above the horizontal meridian has twice the contrast of the one below the meridian (30% vs 15%). The two patches on the horizontal meridian have 10% contrast. It takes a 30% patch in the North to match the 10% patch equidistant from the plus sign in the East. Much of the work of investigating this phenomenon comes from Marisa Carrasco’s lab. See [Cameron et al. \(2002\)](#) and [Carrasco et al. \(2001\)](#). Note that there is a large degree of variation from person to person so the patches may not look exactly the same in contrast to you. (The patches are called “Gabor patches” or sometimes just gabors.) Thanks to Jared Abrams for making this figure for me. @copyright Ned Block

past in visual short term memory for 1-3 seconds, albeit at a slightly lower level ([Montaser-Kouhsari & Carrasco 2009](#)). These differences are thought to be due to anatomical asymmetries ([Abrams et al. 2012](#)).

I will assume that the percepts of North and East have the same contrast phenomenologies when seen (simultaneously) in peripheral vision. Of course the fact that they don’t look different

does not prove that they look the same. And their looking the same does not prove that the phenomenology of each of the two patches is the same— as we know from the phenomenal Sorites problem ([Morrison 2013](#)). However, the fact that they look the same is *evidence* that they are the same phenomenologically and we would need a reason to resist that conclusion. Similar issues will be taken up later in section 8 and 10.

I take it as obvious that the North and East patches are determinately different in apparent contrast when sequentially foveated and attended. The fact that the percepts are sequential makes it unlikely that we are misled about the determinate difference by any analog of the “beats” one hears when guitar strings vibrate at slightly different pitches. (I will return to this issue in section 10.)

North and East look the same in peripheral vision and different in foveal vision. How could this be explained in terms of representational content? The only representational explanation I can think of would be based on the idea that the content of foveal representation of contrast is more precise than the content of peripheral representation of contrast. However, as I will explain below, there is evidence that the representation of contrast in the fovea is the same in precision as the representation of contrast in the periphery. So the burden is on the representationist to explain the difference between foveal and peripheral experience of contrast without appeal to a difference in representational precision. I will now turn to a much longer version of the argument which does not have the form of a burden of proof argument but which makes use of the notion of phenomenal precision.

I claim that when you fixate on and attend to the cross, both your perception of the North patch and your perception of the East patch normally veridically represent the contrasts of those patches despite the fact that one sees them only in peripheral vision. Many details cannot be seen in peripheral vision but what can be seen is seen veridically in normal circumstances. Of course the comparisons are illusory: patches that are different in contrast look the same. But the issue I am raising is whether the individual percepts of single patches are illusory. One reason to think there is no illusion is that the same kind of differences in perception caused by spatial inhomogeneities in the visual field occur in all percepts due to *temporal* inhomogeneities—that is, random noise in the visual system that differs from percept to percept. Any two percepts of the same items at the same point in the visual field with

the same degree of attention are likely to differ in apparent contrast (and other properties) due to these random factors. It is hard to see a rationale for supposing that spatial inhomogeneities engender illusion while claiming the opposite for temporal inhomogeneities. And claiming that both engender illusion would make most perception illusory.

This is where my appeal to Tyler Burge’s recent book comes in (2010). As Burge notes, we can explain the operation of constancy mechanisms in perception only by appeal to their function in veridically representing the distal environment. And that function precludes perception being mostly non-veridical.⁴

I will say more by way of justification of the veridicality claim later but for now let us accept that claim and think about the consequences for representationism. Note that the veridicality assumption is meant to apply to non-categorical perception of properties that admit of degrees and is not meant to apply to categorical perception. Afraz et al. (2010) showed that gender neutral faces are more likely to look male in some areas of the visual field and female in others. The veridical percept in this case would represent the gender-neutral faces as androgynous so both of the percepts described are non-veridical. Many varying mag-

4 The popular “predictive coding” framework (Clark 2013; Hohwy 2013) is a kind of Bayesian approach that is sometimes thought to provide a revolutionary alternative to the view of perception as constitutively involving veridical conditions. Of course all of vision science involves a background of Bayesian probabilistic processes. And prediction in the visual system is ubiquitous and important. But these approaches do not undermine the veridicality of perception. A recent review of Jakob Hohwy’s 2013 book on predictive coding (Wilkinson 2014) singles out the predictive coding explanation of binocular rivalry as the parade case, claiming that the predictive coding framework “provides a very satisfying account of binocular rivalry.” Clark (2013, pp. 184-185) also emphasizes the supposed explanation of binocular rivalry. Binocular rivalry is a surprising visual phenomenon in which different stimuli are presented at the same time to the two eyes, e.g., a face to one eye and a house to the other. What the subject sees however is an alternation between a face filling the whole visual field, then a house, then a face, etc. It is widely agreed in vision science that the rough outline of the binocular rivalry phenomenon is explained by a combination of reciprocal inhibition and adaptation: the competing interpretations reciprocally inhibit one another, and when one is in the ascendancy, adaptation weakens it until the other takes over. Hohwy and his colleagues more or less concede this (Hohwy et al. 2008) saying that the predictive coding framework explains why we have reciprocal inhibition and adaptation in the first place. But to the extent that this reflects what is good about the predictive coding framework, it is not a revolutionary alternative to standard vision science but rather an evolutionary gloss on it.

nitudes such as size and contrast are not perceived categorically in this way so there is no corresponding “reality check” for such magnitudes. (Some magnitudes such as orientation may mix categorical and non-categorical perception.)

A percept that attributes a property to an item is veridical only if the item has the attributed property. However, the veridical percepts of North and East (when fixating the cross) attribute the same contrast property since they look the same in contrast. Let us ask what the content of the (veridical) percepts of North and East are when one is fixating the cross and they look the same in contrast. That is, what contrast would the percepts of North and East attribute to those patches? Since East is a 10% patch and North is a 30% patch, and both are veridical, it follows that the percepts have to attribute the same contrast to them (since they look the same). What attributions would be the same and also veridical? The patches would have to be represented as having a range of contrasts between 10% and 30% at a minimum. That is, the minimal imprecision in the representation is 20%, the imprecision of a representation of 10%-30% contrast (including the endpoints).

Now let us ask what the contrast-content is when we fixate (and attend to) the East patch, the 10% patch. If the precision is the same as in peripheral perception (i.e., 20%), the percept could have a content of 10% plus or minus 10%, i.e., 0% to 20%. (A 0% contrast patch would be invisible, so presumably imprecision ranges should be weighted towards higher absolute values of the magnitude perceived. Variability in perceptual response increases with the absolute value of the magnitude perceived—one form of the Weber-Fechner Law. This is a complication that I will mainly ignore.) And for similar reasons, if the precision is the same in foveal as in peripheral perception, the contrast content of the percept of the North patch when one fixates it would be 20%-40%.

The representational precision is 20% but what about the phenomenal precision? Can we make sense of this idea? As with all that is phenomenal, no definition is possible. The best that

we can do is indicate a phenomenon that the reader has to experience for him or herself. One type of example exploits the difference between an object close up and the same object at a distance. An object may look to have the same properties at both distances but with different precisions. An object may look crimson close up but merely red (and not any particular shade) at a distance.

If the phenomenology of perception is grounded in its representational content and if there is such a thing as phenomenal precision, an increase in phenomenal precision depends on a corresponding increase in representational precision. Representational precision can be indexed numerically—a representational content of the length of something as 1 inch—2 inches (i.e., between 1 and 2 inches) is more precise than a representational content of it as 1 inch—3 inches. According to representationism, phenomenal precision is just the phenomenology of the precision of representational content. We experience a percept with representational content of 1 inch-2 inches as having more (i.e., narrower, smaller range) precision—as being more phenomenally determinate—than we experience a percept with representational content 1 inch—3 inches.⁵

Note that I am not saying that we can always ask whether a certain item of phenomenology is more precise or less precise than a certain representational content (though I think there are some cases where this does make sense). What I am saying is that a representationist has to hold that a difference in phenomenal precision is grounded in a difference—of the appropriate sign and magnitude—of representational precision.

Here is the application of these ideas about precision: Foveate North and East in turn (i.e., serially). I claim that they look determinately different. According to what I mean by looking determinately different, for items to look determinately different, their phenomenolo-

⁵ For a direct realist, phenomenal precision is just the precision of the properties we are directly aware of. We can be directly aware of properties with different precisions, for example, crimson, or alternatively red. Similarly we can be directly aware of a 10%-20% contrast property and also a 10%-30% contrast property and the difference constitutes a phenomenal precision difference.

gies cannot be almost completely overlapping. Why is lack of almost complete overlap important? The representational contents of perception can be very imprecise even though discrimination is fine grained. One might represent one patch as 10%-30% in contrast and another patch as 10.5%-30.5% and as noted by [Jeremy Goodman \(2013\)](#) that would in principle allow for discrimination between them. If the phenomenal precision of these percepts is also very wide, then the phenomenologies of these percepts would not be determinately different from one another—given what I mean by these terms.

Don't get me wrong: I do think that items can look different on the basis of different but overlapping contents. For example, if one is foveating a patch and simultaneously sees a patch of the same contrast in peripheral vision, the two will look different in contrast. Each of the two percepts can be veridical (even though the comparative percept is not). And being veridical and being of the same contrast, the intervallic contents have to overlap.

You may be skeptical about whether there is such a thing as phenomenal precision and whether there is such a thing as phenomenal overlap. But a representationist should not be skeptical. If one's visual experience represents one length as between 1 inch and 2 inches and a second as between 1 inch and 3 inches, then it is hard to see how a representationist could deny that the phenomenal character that is grounded in the first is more precise than the phenomenal character that is grounded in the second. And if one patch is represented as 10%-30% in contrast and another patch as 10.5%-30.5% the representationist would need a good reason to claim that the phenomenologies did not almost completely overlap. Given that representationism would seem to be committed to phenomenal precision and phenomenal overlap, it would seem legitimate to assume them in an argument against representationism.

North and East look the same when fixating the cross and determinately different when fixating (and attending) to each in turn. What does this fact tell us about representational and phenomenal precisions? The phenomenal preci-

sion of perception of contrast must be narrower (i.e., smaller range, greater precision) in foveal vision than in peripheral vision—in order to explain why North and East look the same in respect of contrast in peripheral vision but determinately different in foveal vision. Even if we cannot make sense of an absolute value of phenomenal precision at least we can make sense of differences in it. We might think of this as a phenomenal precision principle:

If two things look the same in peripheral vision and determinately different in foveal vision, then the phenomenal precision of foveal vision is narrower (smaller range) than that of peripheral vision.

At least for one of the foveal percepts, and why would one have narrower precision but not the other? And so according to the representationist, representational precision must be narrower in foveal than in peripheral vision as well—otherwise there would be a difference in phenomenal precision that was not grounded in a difference in representational precision. (The peripheral perceptions are simultaneous and the foveal perceptions are serial. The inhomogeneities described here hold both for simultaneous and serial presentations, albeit at a slightly reduced level in serial presentations. This has been shown separately for inhomogeneities in the visual field ([Montaser-Kouhsari & Carrasco 2009](#)) and for the attentional effects to be discussed later ([Rolfs et al. 2013](#))).

Note that as far as the doctrine of supervenience of phenomenology on representation is concerned, North and East could look the same but still be represented differently. The grounding formulation says: With qualifications to be mentioned: different representational contents require different phenomenologies; supervenience speaks to the converse only. Qualifications: there may be different representational parameters, only one of which is the ground of the relevant phenomenology. So there could be multiple representational realizations of a single type of phenomenal state where the representational differences reflect differences in the parameters that are irrelevant to grounding. And:

phenomenology might be grounded in representational content even though the grain of phenomenology is coarser than that of representational content. So there might be differences in fine-grained representational content that do not make a phenomenal difference.

No one would object to the idea that pure dispositions like fragility or solubility could be grounded in different molecular structures in the case of different substances. And physicalists about phenomenology have held that the underlying basis of a common phenomenology might be one physical state in humans and another in robots. However, I have argued that the grounding framework reveals that physicalists should not acknowledge this kind of multiple realizability (2014a). Applied to this case, the idea is that a representationist account should give us a representationist answer to the question of what it is in virtue of which the phenomenology of the peripheral percepts of North and East are the same. Phenomenal sameness requires representational sameness as a ground. And that representational sameness in this case has to be a precision range of 10%-30% or more.

Of course the notions of phenomenal precision and almost complete overlap of phenomenologies are obscure. The methodological situation we are in is that we have a well-developed science of perception but very little science of the phenomenology of perception. One response—very common until recently—is to avoid issues of phenomenology like the plague. But the time may be ripe to try to leverage the science of perception to get some insight into the phenomenology of perception. And that project cannot help but start with some vague intuitive notions.

Here is where we are: foveal percepts of the contrasts of North and East are determinately different in phenomenology but peripheral percepts of them are the same in phenomenology so the phenomenal precision of North and East, each seen foveally is narrower than the phenomenal precision of North and East seen peripherally. If representationism is to avoid a difference in phenomenal precision that is not based in a corresponding and commensurate dif-

ference (of the right direction) in representational precision, then representational precision has to be narrower in foveal than peripheral perception. That is, the representationist should hold that peripheral perception is less representationally precise than foveal perception.

Here comes the punch line: Robert Hess & David Field (1993) compared the discrimination of the locations and contrasts of patches of different contrasts. They presented triples of patches in which the middle patches could differ from the flankers in (1) locations and (2) contrasts. They asked subjects two questions concerning each triple: whether the middle patch differed from the flanker patches in location and contrast. What they found was that discrimination of locations falls off greatly in the periphery but discrimination of contrasts does not. They conclude (pp. 2664, 2666), “... we show that for normal periphery, elevated spatial uncertainty is not associated with elevated levels of contrast uncertainty at any spatial scale... A change in positional error of a factor of 14... from the fovea to the periphery has an associated contrast error that does not significantly increase over the same range of eccentricities.” The graphs are striking: position error increases greatly with peripherality of the stimuli but contrast error is a flat line. See Figure 3 for one of the figures that illustrates this fact. As far as I can tell, this result is widely accepted. Even a critical reply (Levi & Klein 1996) says “Their results (discussed below) show that position discrimination is selectively degraded in the periphery, while contrast discrimination is not affected.” Levi and Klein dispute the alleged explanation of the result, not the result.

Note that I am taking the fact that contrast discrimination does not diminish in peripheral vision to be evidence that representational precision does not decrease in peripheral vision. Hess & Field (1993) describe a model of the result in terms of constant “uncertainty” for contrast across the visual field but increasing uncertainty for location. Their concern is whether the subjects’ visual representations produce locational errors as a result of “undersampling”. They argue that undersampling should affect contrast errors too.

And because it does not, they conclude that the explanation is “uncalibrated neural disarray”: “We propose that, for reasons as yet unknown, the periphery, unlike the fovea, has not undergone sufficient self-calibration to resolve all of its innate anatomical neuronal disorder...” (p. 2669). But we don’t have to buy into neural disarray to accept the observation that contrast uncertainty does not decrease in the periphery.

We feel that foveal attended perception is “crisp”, i.e., high in precision but for some properties—contrast and probably gap size, spatial frequency (stripe density) and speed—there is some reason to think that foveal and peripheral perception are equally precise. The resolution is that some properties—e.g., location—really are represented more imprecisely in the periphery than in the fovea (by a factor of 14). (And some properties seem to be represented *more* precisely in the periphery, e.g., flicker rate for some spatial frequencies; Strasburger et al. 2011). So we can’t think of peripheral perception as imprecise in regard to all properties we can see. And for the properties that do not decline in precision in the periphery, the representational point of view doesn’t seem to work very well.

Acuity is lower in the periphery than in foveal vision. Anton-Erxleben & Carrasco (2013) describe five mechanisms that jointly explain the decrease in acuity with eccentricity. Cone density and the density of the retinal ganglion cells that process cone signals decrease with eccentricity. In addition, average receptive fields are larger in the periphery. (The receptive field of a neuron is the area of space that a neuron responds to. See glossary.) So the elements of a grid will not be visible in the periphery if they are too finely spaced (i.e., if the spatial frequency is too high). To compensate for this, Hess & Field used only very coarse grids in the periphery. So what the result suggests is that contrast uncertainty does not increase in the periphery—for grids that one can actually see in the periphery.

But why do the behavioral results reflect on representational precision rather than phenomenal precision? The anatomical asymmetries

that are the probable basis of the inhomogeneities discussed here are bound to affect unconscious perception in the same way as conscious perception.

The Hess & Field result shows a kind of homogeneity in the visual field in regard to contrast but as I have emphasized in regard to the phenomenon of Figure 2, the visual field is inhomogeneous in regard to contrast. How are these compatible? The inhomogeneities in Figure 1 reflect contrast sensitivity whereas the homogeneity showed by Hess & Field reflect contrast precision.

Here is the argument summarized:

1. The peripheral percepts of North and East, being the same in contrast phenomenology, are the same in contrast-representational contents—if phenomenology is grounded in representation.
2. The peripheral percepts of North and East are both veridical; that is, North and East have the properties attributed to them in peripheral perception.
3. Given veridicality and the difference between North and East in actual contrast, the representational contents of the peripheral percepts must be rather imprecise. Since North is 30% and East is 10%, and since the content characterizes both, the peripheral representational contrast-content has a precision range of at least 10%-30%.
4. Foveal percepts of North and East—one at a time—are determinately different in phenomenology.
5. The phenomenal precision principle: If two things look the same in peripheral vision and determinately different in foveal vision, then the phenomenal precision of foveal vision is narrower (smaller range) than that of peripheral vision.
6. So the phenomenal precision of the foveal percepts of North and East must be narrower than that of the peripheral percepts of these patches.
7. Representationism requires that a difference in phenomenal precision be grounded in a commensurate difference in representational precision.

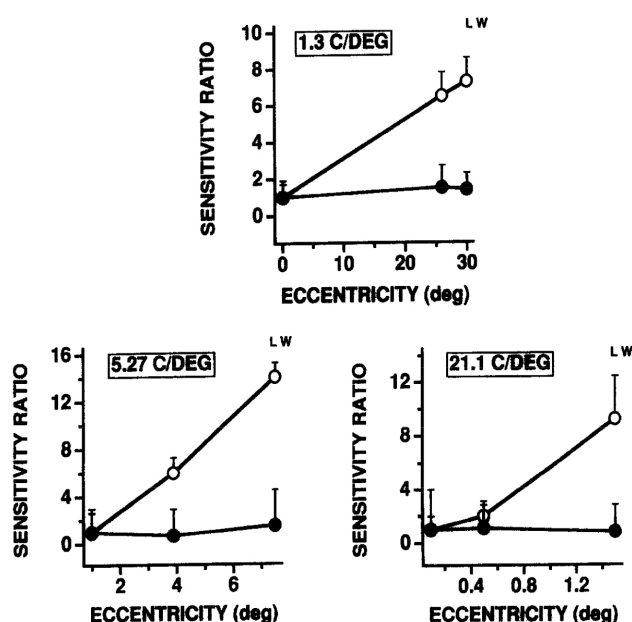


Figure 3: This is one of four graphs from (Hess & Field 1993) showing the comparison between the sensitivity to contrast as compared with the sensitivity to location. The Y-axis represents foveal sensitivity divided by peripheral sensitivity so a value of more than 1 represents greater foveal sensitivity. The solid dots represent contrast sensitivity whereas the open circles represent location sensitivity. The top graph shows sensitivity up to 30 degrees from the line of sight for a very coarse grid of 1.3 cycles per degree. The bottom two graphs show sensitivity for finer grids but at much lower eccentricities. (Coarse grids are visible in the periphery but fine grids would look like a uniform gray surface in the periphery.) Foveal discrimination thresholds are given an arbitrary value of 1. (This is referred to in the article as the values being “normalized”.) What this figure and the other 3 figures show is that contrast sensitivity for grids that are coarse enough to see is the same in the periphery as in the fovea but location sensitivity is much worse in the periphery. Reprinted with permission of *Vision Research*.

8. So representationism requires that the foveal representational precision be narrower than the peripheral representational precision. However the experimental facts suggest maybe not.⁶

⁶ This argument can be stated in direct realist terms but it would require an analog of the veridical/illusory distinction in direct realist terms. See Block (2010) and footnote 14.

9. Conclusion: there is some reason to think that the phenomenology of perception is not grounded in its representational content.
10. The same argument applies to views that hold that there is a kind of “phenomenological representational content” that emanates from the phenomenology of perception (Bayne 2014; Chalmers 2004; Horgan & Tienson 2002). If there were such a thing, it would have to be precise enough to properly reflect phenomenology but imprecise enough to handle the veridicality considerations raised here. And the argument presented here suggests that can’t happen.

The premise that I think needs the most justification is 4. Do we really have enough of a grip on what it is for percepts to be determinately different in phenomenology to justify the idea that the foveal percepts do not have almost completely overlapping phenomenal characters?

Given the problem with 4, I should remind the reader that I started with an argument that did not appeal to phenomenal precision. North and East look the same in peripheral vision and determinately different in foveal vision. How could this be explained in terms of representational content without appealing to a difference in representational precision between fovea and periphery? This argument has the usual problem of a burden of proof argument but it has the advantage of avoiding the obscurity of phenomenal precision.

Another more introspective route to the same conclusion derives from the point mentioned at the beginning that it is natural to feel that the phenomenology of seeing the contrast between lines and spaces foveally differs in precision from seeing the same lines peripherally. The foveal percept seems more “crisp” than the peripheral percept. If this intuitive judgment is correct, there is a discrepancy between the precision of phenomenology and the precision of representational content.

I think this argument gives the reader a pretty good idea of the dialectic of the paper though the paper is more concerned with the issue of change in precision due to differences in attention than with peripherality.

Worth Boone has argued against my point of view using two-point thresholds of tactile discrimination (2013).⁷ As I will explain, I think some of the issues he raises actually support my conclusion.

Boone noted that there are large differences in representational determinacy (precision in my terminology) between tactile acuity as measured by two-point thresholds at various points on the body but that—contrary to what I have said—the precision of the phenomenology matches the precision of the representational content.

First, what are two-point thresholds? “Subjective” two-point thresholds are based on one or two sharp points (e.g., pencil points) being placed at constant separations at various body parts, with subjects reporting whether it feels like there are two points or one point. Objective two-point thresholds are measured by stimulating the skin with either one or two sharp points and observing to what extent the subjects are able to discriminate between these stimuli. Objective thresholds are based on whether there actually are two rather than one point whereas subjective thresholds are based simply on the judgments themselves, independently of their accuracy. The subjective method shows extremely high variability within a single subject on the same body part for a variety of reasons. The objective method has a number of paradoxical features that I won’t go into but if you are interested you can read a short article dramatically titled “The Two-Point Threshold: Not a Measure of Tactile Spatial Resolution” (Craig & Johnson 2000).

However, a recent review (Tong et al. 2013) suggests better measures of tactile acuity that confirm Boone’s point that tactile acuity varies enormously from one part of the skin to another. A glance at a graph in the Tong, et. al. paper reveals that acuity on the tip of the finger is about 5 times that of the palm and about 20 times the acuity on the forearm.

We can ask: is the phenomenology of these perceptions as imprecise as the representational content? Boone says yes but he is judging the

phenomenology of two-point perception. That method is doubly illicit, first because it is unclear that two point discrimination is a measure of anything tactile. Second, the two point judgments may simply reflect the representational contents rather than or in addition to the phenomenology, contaminating the verdict on the very point at issue. If you ask someone how determinate their phenomenology of a two point stimulus is, they may simply be reporting how sure they are they are perceiving two points rather than one. The latter is suggested by considerations of representational “transparency” or “diaphanousness” of experience (Stoljar 2004)⁸. As Thomas Metzinger puts it, we “look through” the experience to its object (2003, p. 173). If so, the phenomenology of judging one vs two may be contaminating the judgment of the precision of the percept.

So I will ask again: Do the differences in phenomenological precision between fingertip and palm and between fingertip and forearm perception differ by factors of 5 and 20? The question is not well formed: we cannot ask about either phenomenological or representational precision without specifying what is being represented.

To get a better question, let us focus on the perception of location. Representational locational imprecision does vary with location on the body. The explanation of the variation is that the number and spatial distribution of sensory receptors that feed into a single sensory neuron (i.e., the receptive field of the sensory neuron) varies widely over the body.⁹ Is there a matching change in the phenomenal precision with regard to location? If there were a massive decrease in representational precision of location from fingertip to forearm without a corresponding decrease in phenomenal precision, we would have a violation of grounding (of a different kind from those already discussed). I suggest you put a single pencil point on your finger tip

⁸ G. E. Moore (1903) famously said “... the moment we try to fix our attention upon consciousness and to see what, distinctly, it is, it seems to vanish: it seems as if we had before us a mere emptiness. When we try to introspect the sensation of blue, all we can see is the blue; the other element is as if it were diaphanous ...”

⁹ For an amusing account of the facts surrounding these issues see Ramachandran & Hirstein (1998).

⁷ I used Figure 2 in a talk at Pittsburgh where Boone was in the audience on November 2, 2012.

and then on your palm and forearm. (Or if you have a helper, do them simultaneously.) If there is a five-fold or twenty-fold difference in phenomenological precision of location it should be appreciable with any stimulus. My own introspective judgment is that there is little or no difference in precision of representation despite the five-fold difference between the fingertip and palm and 20-fold difference between the fingertip and forearm. I am pretty sure that the percepts are not determinately different. No doubt people differ both in these experiences and in their introspective access to them. And with all difficult phenomenal judgments, contamination by theory is no doubt a major source of variability.

If my judgment is right, we have a case of a difference in representational precision without a corresponding difference in phenomenal precision. In the visual case just mentioned, we have evidence for a difference in phenomenal precision without a corresponding difference in representational precision. Taken together, the cases suggest a considerable disconnect between perceptual phenomenology and perceptual representation.

The conclusion of this section is that there is some reason to think that there is no representational content of perception that either grounds or is grounded by the phenomenology of perception—what it is like to perceive.

The reader may wonder how there could be such a disconnect between the phenomenology of perception and its representational content. I mentioned the fact that grip accuracy is about the same in the far periphery (70° off the line of sight) as it is close to the line of sight (5°) despite the fact that conscious vision is extremely weak in the far periphery. Conscious vision is a distinct system from the system that underlies the fine details of perceptually guided action. Though I am not alleging that the system underlying conscious perception is distinct from the system underlying perceptual representation, the upshot of this paper is that they are partially distinct.

In what follows, I will be arguing that facts about attention motivate a similar argument for a discrepancy between the phenomeno-

logy of perception and its representational content. The reason I went through the argument based on inhomogeneities first is that the issues are straightforward compared with the corresponding issues concerning attention. Attention is a complicated phenomenon about which there is a great deal of disagreement, so the rest of the paper has many twists and turns. The argument form as presented so far will not resume until section 8.

4 Attention affects appearance

William James (1890, p. 404) famously said attention “... is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others.” Except for the exclusion of unconscious attention, most scientists would accept something like that characterization today. Spatial attention is attention directed to a portion of environmental space and is distinct from attention to an individual (e.g., a thing, a surface or a property instance) or to a property.

The mechanisms of attention are fairly well understood. Spatial attention boosts neural activation in circuits that process information from the spatial area that is attended, inhibiting activation in circuits that process information from adjacent areas. Feature-based attention boosts neural activation for attended features, inhibiting neural activation for other features. Object based attention does the analogous task for objects. Feature-based attention refines selectivity for the attended feature whereas spatial attention refines selectivity for the attended area of space (Carrasco 2011; Ling et al. 2014).

The main body of this paper is concerned with the effect of the modulation of spatial attention on phenomenology and representational content. Except when mentioned explicitly, I am talking about spatial attention rather than attention to a property instance or an object. My argument is based on experiments that indicate that attention affects appearance. To begin, at-

tention affects perceptual acuity, one measure of which is whether one can detect whether there is a gap or what side it is on in a Landolt square. (For examples of Landolt squares, see Figure 4.)

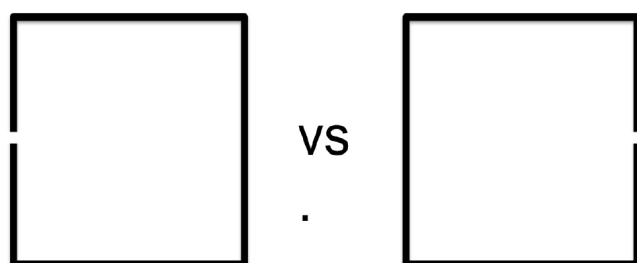


Figure 4: Landolt squares, i.e., squares with gaps. The subjects' task in the experiment diagrammed in Figure 4 was to report (via key presses) whether the gap is on the left or on the right. The squares were presented at various locations while the subject was fixating in the middle of the screen. Redrawn from Yeshurun & Carrasco (1999).

Yaffa Yeshurun & Marisa Carrasco (1999) asked subjects to press different keys depending on whether a Landolt square had a gap on the left or the right. The Landolt square could be presented at any of 16 different locations of 3 different eccentricities. In half of the trials, the square was preceded by a green bar presented briefly at the location in which the square would appear. Then after a pause, a Landolt square appeared in the same location as the line, and then a noise “mask” was presented to prevent an ongoing iconic representation of the stimulus. The subject was supposed to press a key indicating which side the gap was on. See Figure 5 for the sequence of presentations. (An icon would introduce an unwanted source of variability since “iconic memory” varies from person to person. A later experiment (Carrasco et al. 2002) obtained similar results without a mask.) The result is that subjects were more accurate and also faster when the cue indicated the location of the square than when there was no cue. Similar results were shown for other acuity tests, e.g., distinguishing a dotted line from a solid line.

This experiment involves “exogenous” attention in which the subject’s attention is automatically attracted by a highly visible change,

e.g., a sudden motion or disappearance of an object. A similar effect has been shown when the subject is told, for example, to attend to the right when a central bar points in that direction. This is a matter of “endogenous” attention. Exogenous spatial attention is sometimes referred to as “transient” or “bottom-up” attention, whereas endogenous spatial attention is “sustained” or “top-down”. Exogenous attention is involuntary whereas endogenous attention is voluntary. Exogenous spatial attention peaks by 120 ms after the cue, whereas endogenous spatial attention requires at least 300 ms to peak and has no known upper temporal limit.

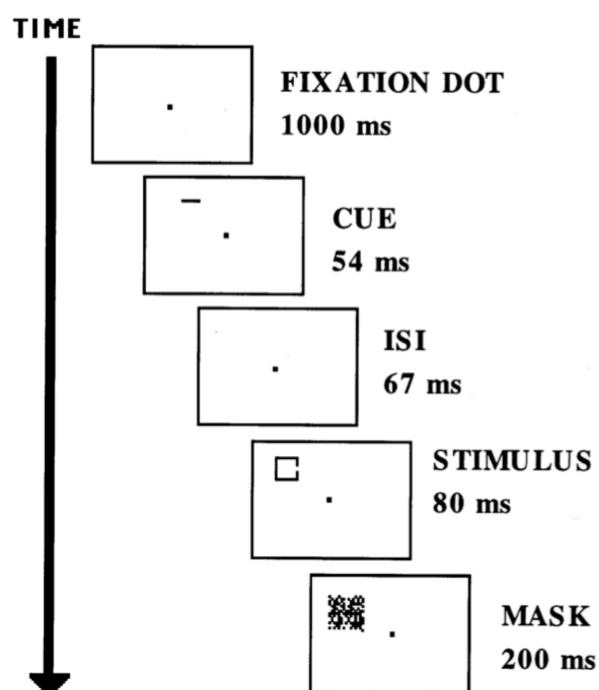


Figure 5: Yeshurun & Carrasco (1999) asked subjects to fixate (point their eyes) at a dot at the center of a screen. Then a cue appeared for 54 ms, then an “inter-stimulus interval”, then a Landolt Square, then a mask. (See the text for the purpose of the mask.) Note that it takes 250 ms for eye movement to a new location, so in this and the other experiments described here the brief presentations of stimuli preclude eye movements to the cued items. I am grateful to Marisa Carrasco for giving me this figure.

Using a similar paradigm and comparing the effects of exogenous and endogenous attention (Montagna et al. 2009), researchers showed that endogenous attention decreased the min-

imum size of a gap that could be detected by about 35% compared to a gap on the opposite side from the cue. That is, subjects could detect much smaller gaps when they attended to the area in which they appeared.

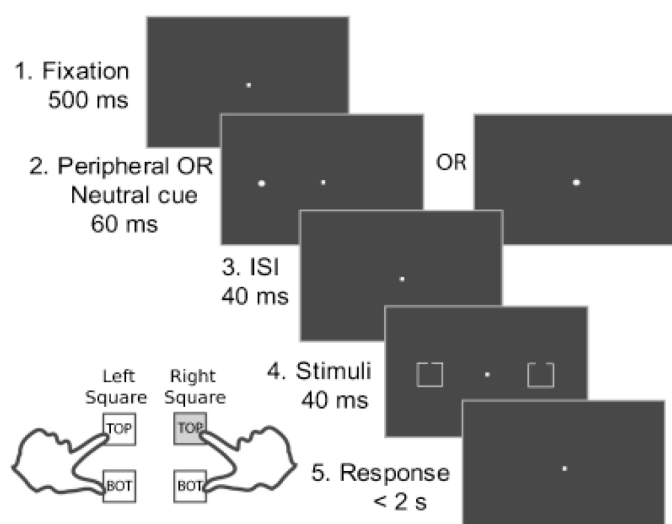


Figure 6: Experiment from [Gobell & Carrasco \(2005\)](#). Procedure described in text. Reproduced with permission from *Psychological Science*.

The conclusion is that attention affects acuity. This is not part of the evidential basis for the argument to come. However there is another effect that is directly relevant to my argument: attention also causes the gap to be perceived as larger. This was shown by a later type of experiment from Carrasco's lab.

The subjects were asked to fixate on the dot that appeared for half a second (upper left in Figure 6). Then the subjects saw a dot on the left or a dot on the right or only at the fixation point in the center of the screen. Then the subjects saw two Landolt squares each of which could have a gap either on the top or the bottom (even though the figure shows the gap on the same side). The subject was then asked to report whether the bigger of the two gaps is on the top or the bottom. If the gap on the left was bigger, the subject was supposed to report the answer using the left pair of keys; *mutatis mutandis* if the right gap is bigger. The subject was told—correctly—that the dot did not predict anything about the size of the gaps. The subjects' instructions focus on the top/bottom

difference whereas what the experiment is really about is the perceived size. The purpose of the dot was to attract the subjects' exogenous (involuntary) attention to one side or the other on some trials. What was being tested is whether attention to, e.g., the left, would cause the perceiver to treat the left gap as bigger. The result was that it does. Subjects did not discriminate between an attended .20° degree gap and an unattended .23° gap. (The gap sizes are measured in degrees of visual angle. If a distance between the eyes and the screen is specified, the degree coding can be changed into inches.)¹⁰ Note that subjects were not asked to judge relative sizes of gaps. In particular they were not asked to judge whether an attended .20° degree gap and an unattended .23° gap “look the same”. Rather, the subjects were asked to make discriminations based on apparent gap size. The result is that they are indiscriminable. And this fact about these experiments has led to disputes about what they really show, as I will explain.

The experiment diagrammed in Figure 5 shows attention increases acuity. This one shows that attention makes gaps look bigger. One of the main mechanisms by which attention improves acuity is that attention shifts and shrinks receptive fields ([Anton-Erxleben & Carrasco 2013](#)). The shifting of receptive fields is probably involved in both the increase in acuity and the larger appearance. Attention to an area of space causes neurons that were not aiming at that area of space to shift towards it. The effect is more neurons covering that area of space. More neurons covering that area increases the acuity of perception of it. And this mechanism is responsible for the increase in apparent size as explained by [Katharina Anton-Erxleben et al. \(2007\)](#). Their explanation depends on the “labeled-line” hypothesis that neurons in the

¹⁰ This effect could be regarded as larger than the result from the Yeshurun and Carrasco paper reported in 5. In that paper, 75% accuracy was achieved by a .20° cued gap as compared with a .22° uncued gap. That difference may be because the 75% accuracy is arbitrary. Or if the difference is real, we could point to the fact that in the Gobell and Carrasco study, the comparison is between an attended square and a square from which attention has been withdrawn, whereas in the Yeshurun and Carrasco study the comparison is between a case in which something is cued and a case in which nothing is cued. Note that there is no need for a mask in this experiment since variations in iconic memory between subjects would be expected to affect equally both the square on the left and on the right.

early visual system are hard wired to code for a certain area of space. So when the receptive fields of neurons shift towards a target the brain treats the size of the target as larger.

5 Is the attentional effect perceptual?

There has been a controversy in the perception literature about whether the kind of effect I have been describing is at least in part genuinely perceptual as opposed to an effect on the decision process involved in generating a report (Schneider & Komlos 2008; Valsecchi et al. 2010).

There probably are effects of attention on aspects of decision, including on conceptualization of a stimulus (Botta et al. 2014). However, I think the case is overwhelming that the attentional effect is at least in part genuinely perceptual. One reason involves “perceptual adaptation” a phenomenon known to Aristotle in the form of the “waterfall illusion”. As Aristotle noted, “...when persons turn away from looking at objects in motion, e.g., rivers, and especially those which flow very rapidly, things really at rest are seen as moving” (1955). Looking at something moving in a direction raises the threshold for seeing motion in that direction, biasing the percept towards motion in the opposite direction.

Perceptual adaptation is involved in the “tilt aftereffect”. If one looks at a left-tilting patch, the neural circuits for the left direction raise their thresholds. This is sometimes described (evocatively but inaccurately—see Anton-Erxleben et al. 2013) as neural fatigue. Then when one looks at a vertical patch, it initially looks tilted to the right. (See Figure 6 of Block 2010). The reason is that the neural circuits for rightward tilt dominate the percept because of the “fatigue” of the leftward tilt neurons. Ling & Carrasco (2006) showed that attending to the adaptor increased the size and duration of a variant of the tilt-aftereffect as if the contrast of the adaptor had itself been raised. Attending to a 70% contrast grating ramped up the tilt after-effect as if the contrast had been raised from 11 to 14% (different magnitudes in different subjects). Ling and Carrasco

directed subjects to attend to gratings for 16 seconds. They found a benefit of attention at first in allowing subjects to distinguish tilts, since attention increases acuity, but then as adaptation increased, discrimination of the adapted tilt was impaired. This kind of adaptation is ubiquitous in perception but does not appear to occur in cognition or decision (Block 2014b). In case anyone thought that the attentional effect was entirely an effect on decision or cognition, this experiment suggests otherwise.

But even apart from the adaptation results, there is strong evidence going back at least to the 1990s from single cell recording in monkeys and in brain imaging for the conclusion that attention increases activity in the neural circuits responsible for the perception of contrast in a manner roughly consonant with an increase in the perception of contrast. Much of this evidence is summarized in sections 4.6 and 4.7 of a review article (Carrasco 2011). My hedge “roughly” stems from debates about the exact effect of attention. There are two kinds of “multiplicative” effects. In “contrast gain” the effect is just as if the contrast of the stimulus has been multiplied by a constant factor. In “response gain” the response is multiplied by a constant factor. The balance of these effects depends on the difference between the size of the target and the size of the “attentional field” (Herrmann et al. 2010). (These ideas are very clearly explained in Chapter 2 of Wu 2014.) A further kind of amplification effect is additive rather than multiplicative: the baseline or “floor” level of activation in the circuit is increased. There is some evidence (Cutrone et al. in press) for increased input baseline as a major part of the attentional effect.

Further, there is plenty of evidence for the conclusion that attention modulates specific cortical circuits depending on what feature is attended. A recent experiment (Emmanouil & Magen 2014; Schoenfeld et al. 2014) compared brain activation when subjects attended to a surface on the basis of its motion and when subjects attended to a surface on the basis of its color. Many of the stimuli involved both color and motion but which feature was task relevant was varied. The result was that motion sensitive

areas of visual cortex were activated first when motion was task relevant and color sensitive areas of visual cortex were activated when color was task relevant. In Carrasco's experiments, subjects' attention is drawn to the specific features that the experiment concerns. In the experiment diagrammed in Figure 6, subjects are directed to report the location of the bigger gap, thereby directing attention to gap size. In the analogous experiment connected with Figure 7, subjects are asked to report the tilt of the patch that is higher in contrast, thereby directing attention to contrast. In experiments concerned with color saturation, subjects are shown stimuli that vary in saturation and asked to report the tilt of the patch that is higher in saturation. Similarly for many other features—speed, spatial frequency, flicker rate, motion coherence, shape, brightness, etc. These instructions can be expected to direct attention to the indicated features with amplification in the circuits that register those features.

Schneider (2011) seems to think that when subjects are asked to report on the side of the larger gap and the gap on the attended side is .20° while the gap on the unattended side is .23°, the subject finds that there is no difference in apparent gap size so the subject just chooses the more salient side. I will discuss salience in the next section, but there is one thing about this charge that raises a distinct issue: that subjects register the increase in apparent size only unconsciously. I now turn to that issue.

6 Is the attentional effect unconscious — like blindsight?

The experiments I have described are “forced choice” experiments in which the subjects must choose between two alternatives. In any perception experiment the issue can be raised of whether the perception is conscious or unconscious, but the issue is often especially troublesome in forced choice experiments with brief stimuli in which subjects make a conscious choice but in which the stimuli are sufficiently evanescent that subjects do not get a really good look at them (Phillips 2011). In addition, the stimuli are presented very briefly, in the ex-

periments described above for 80 ms or less. Many subjects will say that they are never 100% sure of anything. And this can lead to the charge that what is really going on is akin to “blindsight” in which the perception, though genuine, is unconscious (Turatto et al. 2007).

Why are the presentations so brief? Brief presentations preclude eye movements, they preclude significant perceptual adaptation (the “neural fatigue” that causes afterimages), and they preclude certain kinds of strategic responding on the part of subjects. Further, it is known that the effects of exogenous attention peak at around 120 ms after the cue, so to maximize the effects of exogenous attention, brief presentations are required.

Massimo Turatto (2007) showed, using a procedure much like Carrasco's with judgments of perceived speed, that an unattended moving patch was treated as equal in speed to an attended moving patch that was slower by about 10%. However when they asked subjects for subjective judgments of moving stimuli in peripheral vision that really did differ in speed by 10% (without any attentional manipulation), subjects said they saw no difference. Turatto took this as showing that the “just noticeable difference” between the items being distinguished is above the size of the attentional effect so the effect of attention on speed is not conscious. A 10% difference in speed is well above the differences that people can see consciously when they are presented for longer periods, but Turatto argues that for these short presentations the just noticeable difference is larger—that is, it takes a larger difference to be consciously perceived.

There is a difficulty with his experimental procedure though. There are well known problems in asking subjects for same/different judgments. Whether the subjects say ‘same’ or ‘different’ depends not only on their percepts, but also on their decision processes, including how big an apparent difference has to be before they regard it as reflecting reality. These issues are nicely analyzed in (Anton-Erxleben et al. 2010, 2011). When Anton-Erxleben et al. corrected for these deficiencies in another same/different experiment, they found effect sizes that are in

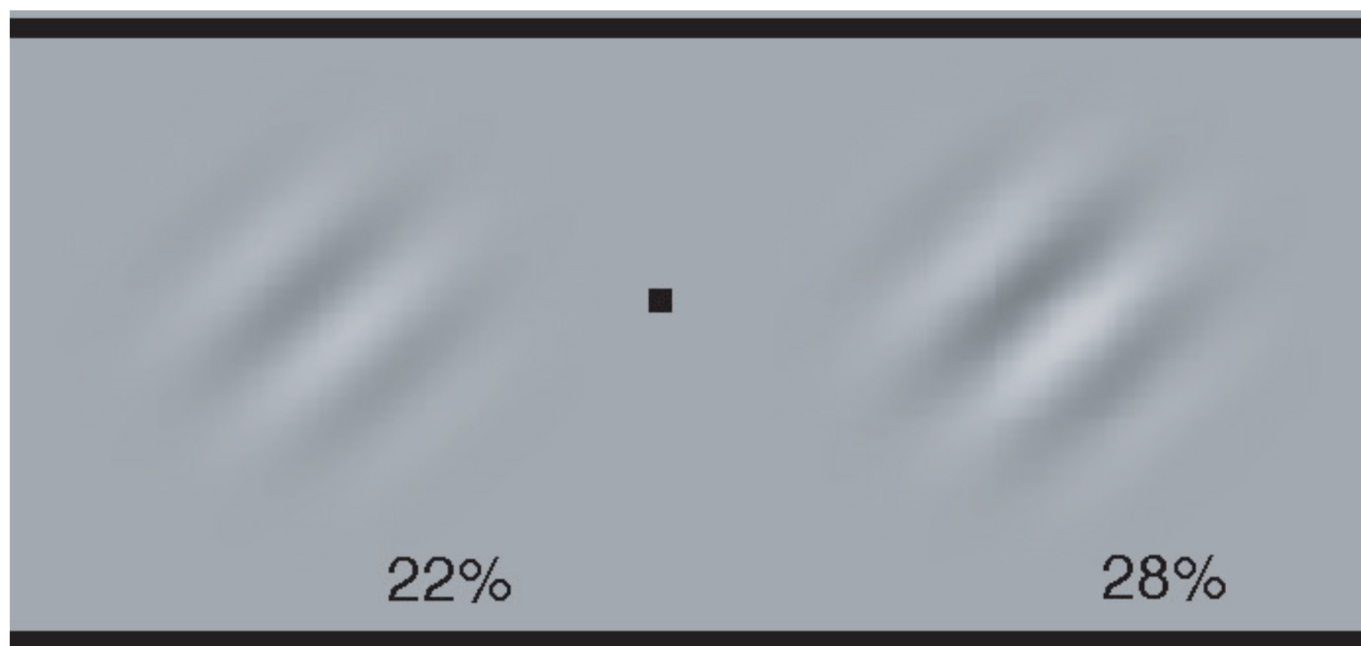


Figure 7: A version of one of the stimuli used in (Carrasco et al. (2004)). Fixate at the dot in the center and move your attention to the left patch without moving your eyes. If you can manage that “covert attention”, the patches should look to have about equal contrast. If you attend to where you are pointing your eyes (the center) you should be able to visually appreciate that the right patch has higher contrast. I am grateful to Marisa Carrasco for supplying this figure.

the vicinity of other paradigms from the Carrasco lab. The effect size is slightly smaller but as they note, that is probably due to inferior sensitivity of the same/different paradigm. (Similar points apply to Kerzel et al. 2010.) One of the conclusions I would draw is that the notion of a “just noticeable difference” in its usual applications is defective because *noticeability is not a perceptual property* but rather the result of an interaction between perception and cognition. I will not go into these issues further here. However, even if Turatto’s methodology is flawed, the issue raised is a good one. How do we know that the effects in Carrasco’s attentional experiments are in fact conscious?

The stimulus in Figure 7 was one of the stimuli used by Carrasco and her colleagues (Carrasco et al. 2004) in the first experiment that demonstrated that attention affects perception by changing the qualities of perception, in this case increasing apparent contrast. The method used was the same as described earlier in connection with Figure 6—in fact this experiment was the model for the experiment of Figure 6. Subjects were asked to report the tilt of the patch that was higher in

contrast after their attention was attracted to one side or the other by a dot as in Gobell & Carrasco (2005). The result was that when the 22% patch was attended it was treated by subjects as the same in contrast as the less attended 28% patch.¹¹ In order to make the judgment, subjects were shown examples of higher and lower contrast. (Contrast is a measure of the difference between light and dark portions of a stimulus.)

As I mentioned, similar experiments have shown that attention increases apparent color saturation, apparent size of a moving pattern, apparent speed, apparent flicker rate, apparent spatial frequency (more about what that is below), apparent motion coherence and apparent time of occurrence—the attended event seems to appear about 40 ms before the unattended event. As I mentioned, the subjects have to take in what parameter the experimenter is talking about—saturation, spatial frequency, gap size, contrast, etc. and then decide which stimulus is greater *with respect to*

¹¹ As mentioned earlier, this methodology tells us that the two were indistinguishable and it is a further step to conclude that they actually look the same.

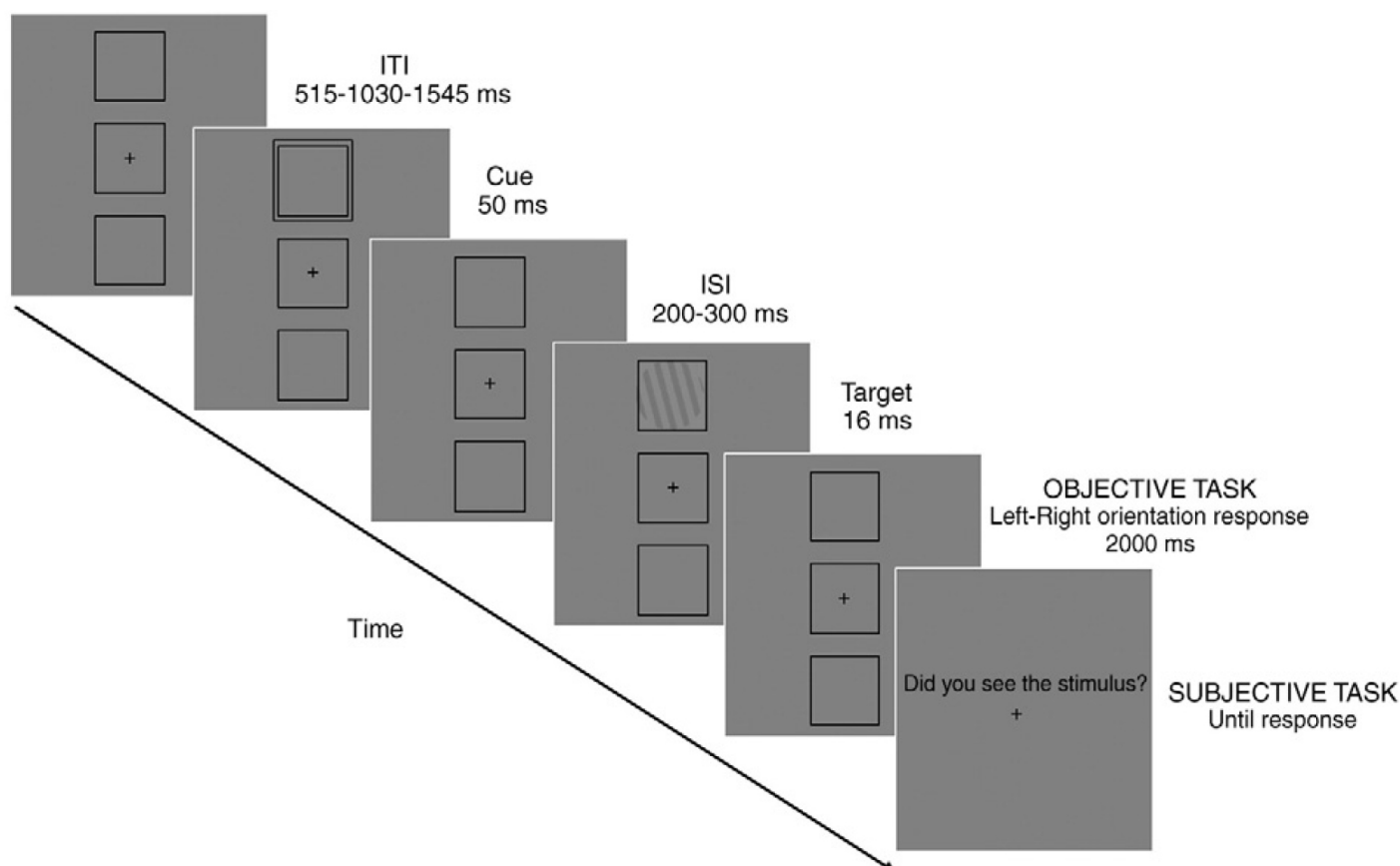


Figure 8: The sequence of events in (Chica et al. 2010) starting from the upper left. ITI = intertrial interval, ISI = interstimulus interval, in this case the period between the offset of the cue and the onset of the stimulus. Reprinted by permission of *NeuroImage*.

that parameter before they can answer the target question about orientation or side that something is on. This is more complex than any certified unconscious perception task that I know of. Further there is positive evidence, summarized in Stanislas Dehaene's recent book on consciousness (2014) that "[m]ulti-step calculations will always require a conscious effort" (p. 95).

What further can be said about whether the effect is conscious? I would be remiss if I did not mention that when you look at a good reproduction of the Carrasco stimuli (Figure 7) you can just see the effect for yourself. (Don't stare for more than a second or two though since adaptation will set in.) It can take a bit of practice to learn to do "covert attention", i.e., to move your attention without moving your eyes though. (In my 2010 paper I included a figure, Figure 2 on p. 32, one of whose purposes was to give the reader practice in covert attention.) Of course you have as much time as you

like to see the effect, whereas in the experiments described you have very little time. Still, what counts for the argument I am making is the effect itself, not its timing. A further difference between just seeing the effect for yourself and the experiments described is that they utilize different types of attention, endogenous for your personal demonstration and exogenous in the experiments described. Endogenous and exogenous attention have been shown to produce roughly comparable effects in Carrasco's experiments, though in some paradigms some exogenous attention is required for endogenous attention to be efficacious (Botta et al. 2014).

I think many people are convinced of the effect because they can just experience for themselves. Not everyone can though as with almost any visual phenomenon. Of course we all know the dangers of relying too heavily on introspective judgments since they are easily manipulated. There is a line of experimentation that addresses part of the issue.

Ana Chica and her colleagues (Chica et al. 2011; Chica et al. 2010) have done a series of experiments that directly address visibility.

The Chica et al. experiment (the 2010 version) presents subjects with tilted patches that are designed to be on the threshold of conscious perception and subjects were explicitly asked whether they saw the target (after making an orientation judgment). Subjects were strongly encouraged to be conservative in saying they saw the target. They were supposed to avoid “false alarms”, i.e., saying there was a target when there was no target, and they saw periodic messages indicating how well they had been doing in avoiding false alarms. In 25% of the trials there was no target.

First subjects saw a fixation point inside the middle of 3 boxes (pictured on the upper left side of Figure 8), then there was a brief cue consisting of a square around one of the boxes. Then the target—a patch oriented either to the right or the left—could appear for 16 ms (even briefer than in Carrasco’s experiments). Next, subjects had to indicate by pressing keys—within 2 seconds—which way the patch was oriented. They had to choose one of the keys whether they saw something or not, i.e., this was a “forced choice” experiment. Then they indicated whether they saw the target or not. The experimenters adopted a procedure—tailored to each subject’s perceptual abilities—to make sure the target was at the threshold of visibility—for that subject. They started each subject with a patch of sufficiently high contrast to see the stimulus. Every 16 trials they lowered the contrast until the subject was not detecting at least 25% of the patches (by the “Did you see the stimulus” test). If the percentage of avowedly seen patches went below 60%, they increased the contrast.

The main result was that the proportion of avowedly seen patches was much higher for “validly cued targets,” i.e., when the cue was on the box that had the patch than when the cue was invalid, i.e., on the box on the opposite side or neutral (when the cue was on the middle box where no target ever appeared). In addition, the reaction time for the cued patches was much shorter than for uncued patches. Chica et al.

also collected brain imaging data that suggested unsurprisingly that the valid cues attracted attention to the cued side, and more interestingly, that when the subjects saw the patch despite invalid cuing (i.e., the cue was on the opposite side), the cue had often failed to attract attention.

This experiment suggests that attention can affect whether a target is consciously visible or not. The subjects were not probed, however, on the issue of whether they actually made their judgments on the basis of the consciously visible tilt. However, when *subjects reported not seeing the target, they were at chance on reporting the tilt. And when subjects reported seeing the tilt, they were substantially above chance.* This is not the profile one sees in blindsight or in unconscious priming where subjects report not seeing the stimulus at all; but more significantly the tight relationship between consciously seeing the stimulus and being able to judge the tilt does suggest that they were reporting the tilt on the basis of the conscious perception.

Chica’s experiments are relevant to the consciousness of Carrasco’s stimuli in another way. Chica’s stimuli were presented very briefly: 16 ms in the experiment just described. Carrasco’s stimuli were presented for longer, up to 100 ms in some experiments. In addition, Chica’s contrasts were very low, as befits stimuli that were supposed to be at the threshold of visibility. The experiment described above does not report contrasts but in other papers with somewhat more complex experiments along the same lines (Botta et al. 2014; Chica et al. 2013), the contrasts required for 50% detection were about 3%; high detection seems to require up to 10% contrast. In Carrasco’s experiments, much higher contrasts are almost always used. I conclude that the reasons against the “blindsight” analogy in Chica’s experiments apply even more strongly to Carrasco’s methodology.

Given the high rates of conscious vision of 16 ms stimulus presentation even at lower contrasts than most of those used in Carrasco’s experiments, I will ignore the issue of brevity of stimulus presentations in the discussion to follow.

Keith Schneider (2011, 2006; Schneider & Komlos 2008) has argued that Carrasco's results are based on salience rather than perceptual variables such as perceived contrast, gap size, flicker rate, spatial frequency, etc. (Recently, Schneider and Jake Beck have written a draft of a paper on this topic. Rather than ascribe any specific view to a paper in draft, I will discuss the issue of salience—stimulated by their remarks—but from my own point of view.) I believe that the Carrasco lab is correct in their experimental and methodological disagreement with Schneider (Anton-Erxleben et al. 2010, 2011), however it would be digressive for me to discuss the issues involved in any detail here. I believe though that it is possible to get some insight without going into those issues.

A crude version of a salience objection treats salience as a “response bias” in the sense of a behavioral disposition to respond (in the basic Carrasco paradigm illustrated in Figure 7) by choosing the attended item. The idea is that when faced with a choice between gaps, the subject is disposed to choose not the gap that looks larger but rather the attended gap. This account is ruled out by a control in many of the Carrasco experiments of asking the subject to report the properties of the smaller gap or the patch that is lower in contrast. The attended side is still boosted in apparent contrast or gap size though the effect can be slightly smaller in magnitude so there is a small effect of “response bias” together with a main effect on perception. Carrasco also showed that choosing the lower contrast patch or smaller gap did not take any extra time, ruling out a version of the behavioral disposition objection that adds on an “inversion of response”.

A more sophisticated salience objection alleges a “decision bias” in the sense of a post-perceptual feature of the cognitive process involved in making a decision of how to respond. All such accounts that I know of are ruled out by the fact, mentioned above, that the effect is substantially perceptual in nature. In addition, Carrasco showed that the effect works for some properties but not others (Fuller & Carrasco 2006). As I have mentioned a number of times, it works for saturation but not hue. Both exper-

iments involved a procedure like that in Figure 6. In the saturation version, subjects were asked to report the tilt of the more “colorful” stimulus, where the stimuli differed in color saturation. In the hue version, subjects were asked to report the tilt of the “more bluish” stimulus, where stimuli differed along a blue/purple continuum. The result: there is an attentional effect on saturation but not hue. A cognitive decision bias should equally affect both saturation and hue. If the subjects are not aware of any difference in hue between the attended and unattended sides, it would seem that the “salience” perspective would say they would choose the attended side. But they don't. Another possibility is that the bias is perceptual in some way, say a matter of perceptual prediction (Hohwy 2013). In either case, the conclusion is that the effect is substantially perceptual and cannot be due simply to any kind of a cognitive decision bias toward choosing attended stimuli.

Whatever understanding of salience the salience objection appeals to, salience must be or be associated with a perceptual property, i.e., a property that is genuinely represented in vision. Some say (Prinz 2012) that the perceptual properties that are involved in vision are limited to a small set whose basic low level representations are products of sensory transduction: shape, spatial relations (including position and size), geometrical motion, texture, brightness, contrast and color. In other words, according to this “lean” theory of perceptual properties, though we speak loosely of seeing something as a face or as a case of causation, in reality seeing-as is limited to a small list of properties that are the output of peripheral sense organs. Others (Block 2014b; Siegel 2010) argue for a more expansive list of genuinely perceptual properties.

How do we know which properties are perceptual? We know that contrast, size, speed, spatial frequency (roughly stripe density), etc. are perceptual properties because they participate in perceptual phenomena, for example in perceptual adaptation and perceptual popout. I mentioned the waterfall illusion in which staring at a moving stimulus makes a stationary item seem to move in the opposite direction. And I'm

sure every reader is familiar with color afterimages. Adaptation is a ubiquitous perceptual phenomenon that can be used to show that size, speed, stripiness, etc. are perceptual properties. Note that I am not trying to *define* the notion of a perceptual property in terms of ...perception. The point rather is that perception is a natural kind and the perceptual nature of a representation is revealed in participating in phenomena in that kind (Block 2014b). By these tests, for example, there is evidence that certain face and facial emotion-related properties are perceptual. Viewers seeing an array of objects including one face can pick out the face very quickly on the basis of “parallel search”, just as they can pick out a red object in a sea of green objects. Similarly there are many adaptation effects for faces and facial expressions.

Is salience a perceptual property in this sense? Attention is important to both cognition and perception, but attention can be perceptual. In order to explain the effect of attention on increasing the duration and magnitude of the tilt after-effect (and the improvement followed by impairment in discrimination) as described earlier, the visual system would have to track or register attention or where or what one is attending to in addition to being affected by attention. As I will explain in the next section, there is an open question of whether the visual system does much by way of tracking attention.

In discussions of salience there is often a conflation between salience as a perceptual property and the genuine perceptual properties that are involved in attracting attention, like high contrast or speed or sudden changes in position. We commonly speak of a saliency map in the sense of the map of locations in the visible environmental layout in terms of whether they are likely to attract attention. The perceptual properties here do not involve salience itself but rather differences with nearby locations in visible feature dimensions (Itti & Koch 2000), for example in visible motion or appearance or disappearance. People also speak of a saliency map in the brain, meaning the increased activations that correspond to attended areas. If salience is supposed to be something other than attention itself, that is If Beck and Schneider are giving

an explanation that is a genuine alternative to Carrasco’s, they have to show that salience is the kind of perceptual property that is registered in the visual system and that can combine in an additive fashion with contrast to affect adaptation as in the tilt after-effect. I know of absolutely no evidence for such a thing.

Many sources of evidence contribute to our knowledge of the fact that attention increases apparent contrast. John Reynolds et al. have shown that attention boosts responses in individual neurons in monkeys (2000). They developed a model of the mechanisms of this and more complex effects involving multiple stimuli (Reynolds & Chelazzi 2004). Many brain-imaging studies have shown similar effects. See sections 4.6 and 4.7 of Carrasco (2011) where much of this work is summarized. At the behavioral level, attention increases sensitivity roughly as if contrast were increased and similarly, attention can mimic the effects of increased contrast on making a stimulus visible (as in the Chica experiment mentioned earlier). And as mentioned earlier, attention increases adaptational effects as if contrast were increased. Every result involving “salience” that I have seen is just a redescription of effects of the sort mentioned.

Here is a way of seeing the emptiness of appeals to salience: As mentioned earlier, at the neural level, there are two main types of response functions by which attention increases the firing rate of neurons, multiplicative and additive. I mentioned a recent paper that compares simulations of neural responses of these sorts to behavioral data in order to ascertain which of the types of amplification are mainly being used by the visual system. Though the multiplicative models work pretty well, one additive model works very well. Thus we have strong evidence for the functional relation between attention and the increase in contrast responses in the visual system. For simplicity, let us focus on the multiplicative mechanisms: In multiplicative gain, the response of the neuron is multiplied by a constant factor. For example, a neuron that responds to orientation will give a large response to its preferred orientation and a smaller response to other orienta-

tions to the extent that they are distant from the preferred orientation. (For example, a neuron that likes vertical lines will give a large response to vertical lines.) Since multiplying a larger number by a constant produces a larger effect, multiplicative gain is most effective for the preferred orientation. A second multiplicative response function is response gain in which the sensitivity of the neuron is multiplied by a constant factor. The effect is one of ratcheting up the response to stimuli of every orientation. The widely accepted normalization model of attention (Reynolds & Heeger 2009) explains the balance of these two mechanisms (and of additive gain) in terms of factors such as the relative size of the target and the attentional field. The attentional function of a given neuron can show more multiplicative gain or more response gain or more additive gain depending on these factors. Here is my point. We can answer the question of what the difference between these response functions is with respect to increasing apparent contrast. For example, multiplicative gain increases the apparent contrast more for the preferred orientation and response gain increases apparent contrast more for unpreferred orientations. What is the answer to the corresponding question for salience? Does multiplicative gain increase salience more for preferred or unpreferred orientations? Is it the same as for contrast? If so, then maybe “salience” is being used as a synonym for contrast. A different answer would be: multiplicative gain and response gain are equally mechanisms of salience. In this latter case it looks as if “salience” is just being used to mean attention. To repeat the general point: Those who advocate a “salience” explanation of the phenomena have to show that there is a property that is (1) perceptual, (2) not contrast and (3) acts in the ways indicated above.

Sometimes the issue is put in terms of “phenomenal salience” (Wu 2014). I think this way of talking just muddies the waters. Perceptual properties can operate in both conscious and unconscious perception. (At least: it would be an amazing discovery that there is a perceptual property that only appears at the conscious level.) Attention—at least exogenous attention—operates in unconscious perception in a sim-

ilar manner to conscious perception (Chica et al. 2011; Kentridge et al. 2008; Norman et al. 2013). Further, it has recently been discovered using optogenetic methods that top-down activation of visual area V1 is about the same in awake and anesthetized mice (Zhang et al. 2014). This top-down activation involved feedback from a brain area in the mouse that corresponds to a locus of voluntary attention in humans. If salience is a perceptual property, it should be operative in unconscious perception. So the salience issue is an issue about perception, not about just conscious perception.

The upshot is that it is not at all clear how a salience objection would work, so the burden is on those who advocate it to explain it. I raised the issue of whether we are aware of where we are attending in connection with whether we are aware of salience, so I now turn briefly to that question.

7 Are we aware of where we are and are not attending?

There are a number of ways of approaching this issue, none of them very satisfying. We are certainly aware of some aspects of voluntary attention—when we “pay attention” to one thing rather than another (“endogenous attention”). But much of attention is involuntary (“exogenous”). Any perceptibly sudden movement, appearance or disappearance or sound will be likely to attract exogenous attention. Subjects in perceptual experiments can try to ignore sudden movements and sounds but they attract exogenous attention nonetheless. Exogenous attention ramps up more quickly (120 ms vs 300 ms) and dies off more quickly. Eye movements can also be voluntary or involuntary. Awareness of where the eyes are pointing is a rough index of awareness of attention. There is some evidence that people are not very aware of the time and direction of their “saccades”, the quick ballistic eye movements that occur when we are visually exploring our environment. Heiner Deubel et al. showed that subjects seem to “have no explicit knowledge about their...eye position” and often don’t “notice the occurrence of even large saccadic eye movements” (1999, p. 68).

However, this is not conclusive evidence that they don't know where they are attending since they may confuse movement of attention with movements of the eyes. And the visual system could track attention even if subjects are not aware of where they are attending.

Perhaps more illumination can be achieved from work on the “landscape of attention” (Datta & DeYoe 2009). Brain imaging shows a complex rapidly shifting map of spatial attention in the visual system. Spatial attention can be “focused” at one location even though there is almost as much attention at a number of other locations and some attention throughout half the visual field. The attentional field often has a “Mexican hat” shape with amplification at the center surrounded by a ring of inhibition and then an increase outside that ring. Certainly no one is aware of all that dynamic detail though I have been unable to find any specific study addressing the issue. I think it is safe to say that in normal perception there is no phenomenology that specifies much of the detail of where one is and is not attending—nor how much one is attending. So any attempt to explain Carrasco's results that appeals to our awareness of where we are attending takes on the burden of showing that we do have sufficiently fine-grained awareness of where we are attending.

8 Veridicality and representationism

In Carrasco's experiments, an attended $.20^\circ$ gap is not discriminated from an “unattended” $.23^\circ$ gap. I think the best conclusion is that attention changes perceived size and contrast. (Recall that I am talking about spatial attention rather than attention to a property instance or an object.) Do the gaps just fail to look different or do they look the same?

In Carrasco's main paradigm, subjects are forced to choose which stimulus is bigger (or faster or higher in contrast). In the case of an attended $.20^\circ$ gap as compared with an “unattended” $.23^\circ$ gap, subjects are as likely to choose one option as the other. In this sense these options are not discriminable. However, I mentioned that when subjects are asked instead

whether the items are the same or different, the effect of attention is slightly smaller. And that may suggest that there is substantive daylight between not looking different and looking the same. (Of course this difference matters very much in some contexts, for example, as mentioned earlier, the context of the phenomenal Sorites issue; Morrison 2013.) As I mentioned earlier; Anton-Erxleben et al. and her colleagues argue persuasively that the smaller effect is due to the same/different paradigm being a less sensitive measure (2011). In what follows I will assume that the attended $.20^\circ$ gap looks the same in respect of size as the “unattended” $.23^\circ$ gap.

I put the “unattended” in quotes because I mean no commitment to the improbable claim that there is no attention on the $.23^\circ$ gap. There is no agreement on whether there can be conscious perception or even unconscious perception with zero spatial attention or whether zero spatial attention is even possible.¹² Indeterminacy in our concept of attention may even make this an unanswerable question. Still, I will adopt the abbreviatory convention of referring to stimuli that are not focally attended as “unattended”.

Since the apparent size of the gap differs depending on where one is attending, the question arises as to which of these various percepts of the gap gets its size right (or most nearly right) and which gets its size wrong (or most nearly wrong). Veridicality is a matter of getting things right and veridicality in perception is a matter of the world being as it appears to be. There would be no good reason to decide that the veridical percept of the gap is one in which one is attending to a spot one inch away from it; why pick one inch rather than one centimeter? (Recall that the attentional landscape of amplification and inhibition varies from place to place and from moment to moment.) The most obvious candidate for a non-arbitrary answer to the question is: the veridical percept of the gap (if there is any veridical percept) is the one in which one is attending to the gap itself.

¹² Spatial attention does not require feature-based attention or attention to an object. See Wayne Wu's book on some of these issues (2014).

I think of veridicality as all or none, but for the sake of accommodating different opinions I can countenance degrees of veridicality. One innocuous use of such a phrase is that if one represents a $.19^\circ$ gap as $.20^\circ$, the percept is more veridical than if one represents it as $.21^\circ$. Also we could say that other things equal, a percept that attributes a higher probability of being $.21^\circ$ to a $.21^\circ$ gap is more veridical.

But once it is stated that the most veridical percept of the gap is one in which one is attending to the gap, one wonders why one should believe this hypothesis rather than the *opposite* hypothesis that attention distorts by magnifying, illusorily, for the purpose of getting information and that the attended item is seen illusorily. Is the perception of the gap with less attention really illusory in the sense of a discrepancy between stimulus and perception?

In an article on Carrasco's discovery, [Stefan Treue \(2004, p. 436\)](#) says this:

In summary, this study provides convincing support for an attentional enhancement of stimulus appearance. It completes a triangle of converging evidence from electrophysiology, functional brain imaging and now psychophysical findings, which argues that attention not only enhances the processing of attended sensory information but manipulates its very appearance. ...attention turns out to be another tool at the visual system's disposal to provide an organism with an optimized representation of the sensory input that emphasizes relevant details, *even at the expense of a faithful representation of the sensory input.* (italics added)¹³

I quote Treue not because I agree with him but in order to get a statement of that view on the table. There is no sufficient reason to accept the view that an attended perception of a gap allows us to see it as it really is rather than the view that attention in perception is like a magnifying glass, distorting for informational pur-

poses at the cost of illusion. I can imagine considerations that might incline one towards adopting one or the other of these positions—that attention falsifies or that attention “veridicalizes”—but the adoption would be for purposes of one or another kind of utility, *not as a principled reason to think that the highest degree of veridicality is really to be found in that case.*

The challenge is to find a principled reason for regarding seeing a thing or place with a certain degree of attention to be more veridical than seeing it with a different degree of attention. Sufficiently decreasing attention to something can move the perception below the threshold of visibility. But not seeing something that is too small to see or to faint to see need not be a matter of illusion.

Chris Hill (in conversation) and [Sebastian Watzl \(forthcoming\)](#) have argued that there is an optimal level of attention and perception with all other values engender illusion. Watzl's version of this view appeals to the idea that the function of attention is to make perceptual representations usable—as opposed to the function of perception of veridically representing the world. These functions will conflict in normal circumstances. The optimal level of attention for fulfilling the function of perception—veridicality—will be achieved in an idealized scenario of no attention, or one of equal attention to everything. This is an interesting point of view, but is contradicted by the point made earlier that veridicality conditions require a history of veridical representation.

Epistemicists about vagueness think that there can be an unknowable fact of the matter as to a sharp border between bald and not bald, a number of hairs that a bald man can have even though adding a single hair will make the man not bald. But if there can be a fact about a border even though there could be no principled reason to regard any particular border as the real one, why can't there be a fact about what degree of attention engenders veridicality that no one could have a principled reason to accept? Epistemicists should not regard the cases as analogous since they think there is a principled reason to hold there is a fact about a border and a principled

¹³ Carrasco ([Carrasco et al. 2008, p. 1162](#)) has been interpreted as agreeing with Treue by [Stazicker \(2011a\)](#) and [Watzl \(forthcoming\)](#). Carrasco tells me she did not mean to endorse the Treue view.

explanation for our ignorance (Sorenson 2013).

It may be objected that there is good reason to accept Treue's point of view since after all, attention to the .20° gap makes it look, illusorily, to be the same size as the unattended .23° gap. But why not blame the illusion on the percept of the unattended gap rather than the attended gap? One can blame the mismatch, but that does not help in deciding whether attention to an individual item engenders veridicality or illusion. I think the issues are clearer when one avoids the comparative perception and just asks, say of the situation in Figure 5, whether perception of the gap can be veridical when it is cued and one is attending to it or when it is not cued and one is attending elsewhere. There is no adequate justification for one answer over the other. Some may wish to abandon the notion of veridicality as applied to perception but that would be to abandon the notion of representational content as applied to perception and so to abandon representationism. The representational content of a perception is—constitutively—the veridicality conditions. There is a strong a priori case for perceptual representation (Siegel 2010). And in any case the science of perception makes essential use of veridicality (Burge 2010).

In the discussion of the analogous issue with regard to inhomogeneities in the visual field, I noted that the sort of differences in perception caused by spatial inhomogeneities are paralleled by differences due to temporal inhomogeneities—that is variation from percept to percept due to random factors. Any two percepts of the same items at the same point in the visual field with the same degree of attention are likely to differ in apparent contrast (and other properties) due to these random factors. It is hard to see a rationale for treating spatial inhomogeneities differently from temporal inhomogeneities and it is hard to see a rationale for treating either of them differently from the inhomogeneities due to distribution of attention. Claiming that all engender illusion would make most perception illusory.

We are considering the question of whether the veridical percept of the gap is the

attended one or the unattended one. But is there a well formed question here? Is it endogenous attention that counts? Or exogenous attention? We are talking about spatial attention but what if feature-based or object based attention goes counter to spatial attention? That is, one can be attending to a place but also to a property that is instantiated in another place. And is it the absolute or relative value of spatial attention that matters? That is, is it some absolute attentional value or is it the most attended place that is seen veridically? Talking on a fake cell phone drains away spatial attention, causing the subjects to miss seeing objects in the centers of their visual fields (Scholl et al. 2003). (Scholl et al. used a fake cell phone to avoid the unnecessary source of variability of features of the responses from the other end of the line.) If it is absolute value that counts then when talking on a fake cell phone (and presumably a real one too), all vision would be illusory. That is a conclusion that we would have to have some very good reason to accept.¹⁴

One caution: I am speaking oversimply in a number of respects in asking whether attention engenders veridicality or illusion. I mentioned the issue of whether veridicality is graded. And there is an independent issue of relativity to what property one is talking about. In the experiment pictured in Figure 5,

¹⁴ It may be thought that the issue of which percept is veridical is avoided by forms of direct realism that hold that there are no perceptual illusions. For example, Bill Brewer holds that in the Müller-Lyer illusion (so called) in which lines of the same length look to be of different lengths, what one is seeing is a resemblance between the situation in front of one's eyes and what he calls a paradigm of different lengths. The idea is that that is what equal lines look like when surrounded by opposite-facing arrowheads. And the way equal lines look in that circumstance is like pairs of unequal lines one has seen. On this form of direct realism, the "illusion" to the extent that one can speak of such a thing is in the mistaken inference that the lines in front of one's eyes are of different lengths. They resemble pairs of lines of unequal lines but one should not conclude that they are unequal.

However, Brewer requires differentiating between cases in which one sees a property instantiated before one's eyes that is not a resemblance to something unseen and the cases in which one sees a resemblance. In effect, the cases in which one sees a resemblance to something unseen is a pseudo-illusion category that he has to recognize. So the question arises of whether this pseudo-illusion arises in the case of attention or in the case of the lack of it. That is, is one seeing a resemblance to a non-existent thing when one attends or when one does not attend? And this is an unanswerable question for the reasons explored in this section.

the attended percept is certainly more likely to be veridical in respect of which side of the square the gap is on. And in the experiment pictured in Figure 6, the comparative percept—that is, the percept of the comparative size between the right and the left is distorted by attention to one side and improved by attention to the fixation point. So veridicality is certainly affected by attention—though in different ways for different properties. The question I am asking about gap size is whether a single gap is perceived more—or on the contrary, less—veridically if it is attended. More generally, there are certain properties—I have mentioned size, contrast, color saturation and others—for which attention to an individual item changes appearance of that property. Which is more veridical, the pre-change or post-change appearance?

One might think that there is a simple way to get at the issue of whether attention magnifies, illusorily, for the purposes of getting information, or whether attention makes things look more as they really are. You could just ask people how contrasty a patch is or how big a gap is and then consider whether those answers correspond better to reality when perception is attentive or inattentive. But the human ability to make such absolute judgments for at least some relevant dimensions is remarkably poor, certainly orders of magnitude worse than our ability to discriminate stimuli (Chirimuuta & Tolhurst 2005a). In particular, the uncertainty of absolute identification (absolute in the sense of the ability to say what the contrast is in percentage terms) is far larger than the effects of attention. Even if there were some sort of statistical advantage or disadvantage to conditions of attention in estimating contrast or gap size one would have to ask whether the advantage could be ascribed to better perception or to better inference from a percept that did not differ in veridicality.

I will assume in what follows that attended and unattended perception can both be veridical. Considerations of the same sort mentioned here also apply to the veridicality of perception in both the upper and lower visual field.

9 Indeterminate contents and the phenomenal precision principle

As I mentioned, an attended $.20^\circ$ gap looks the same size as an unattended $.23^\circ$ gap. Of course the comparative percept—the gaps looking the same—is illusory. But what about the percepts of each gap, considered separately? Is the percept of the attended $.20^\circ$ gap illusory? Is the percept of the unattended $.23^\circ$ gap illusory? I argued that we would need a better reason than we have to suppose that one but not the other is illusory. And I claimed that we should not suppose that both are illusory. The option I have argued for is that both are (or rather can be in normal circumstances) veridical. As I mentioned, the simplest perceptual representations contain two elements, a singular element that represents an individual item and a perceptual “attributive” in Burge’s terminology that attributes a property to that individual item (2010). In the gap-size case, the perceptual attributive attributes sizes to gaps. A veridical percept attributes a size to a gap only if the gap has that size. In order for the attributed property to apply to both gaps, that property will have to be “intervalic”, i.e., a somewhat imprecise property—for example, the property of being within the range of $.20^\circ$ to $.23^\circ$ (inclusive of endpoints). Since both gaps are in that range, both percepts are veridical (in respect of gap size).

Perhaps a probabilistic treatment of these ranges of values is in order? But how can one justify one probability distribution rather than another without making assumptions about whether the attended gap is seen more veridically than the unattended gap? For example, to say that the unattended percept of the $.nn^\circ$ gap represents the gap as most likely to be $.nn^\circ$, whereas the attended percept represents the same gap as most likely to have some other value is to regard the unattended percept as more veridical than the attended percept. A probabilistic treatment would perhaps pass the sufficient reason test though if the same probability were attributed to both ends of the range.

It will be useful to move back to the example of contrast. The data portrayed in Figure

7 comes from the bottom right of Figure 9. Figure 9 contains four comparisons, each of which is keyed to one of the four little squares between the patches. If one fixates on one of the squares, the patch to the left of the square attended is the same in apparent contrast as the patch on the right unattended. The 22% patch can be unattended—in which case it has the same apparent contrast as the 16% patch when it is attended, or the 22% patch can be attended in which case it has the same apparent contrast as the 28% patch when it is unattended. So different veridical percepts of the 22% patch could represent it as the same as patches that are 6% more or 6% less in contrast.

Consider the contrast phenomenology of an attended percept of the 22% patch. That phenomenology is the same as the phenomenology of a 28% patch unattended. Assuming that there is not normally a phenomenology that specifies what one is and is not attending to, a matter discussed above in section 7—the phenomenology of a 22% patch attended does not carry the information of whether it is the phenomenology of a percept of a 22% patch or of a 28% patch. So in order for both percepts with that phenomenology to be veridical, the representational content would have to be at a minimum 22%-28% (inclusive of 22% and 28%).

However, there is a determinate difference in phenomenology between percepts of the 22% patch and the 28% patch when serially fixated and attended as you can verify by looking at Figure 9. (There are larger differences of this sort to be described later and as I mentioned in section 3, inhomogeneities in the visual field produce larger differences of this sort.) I believe that this determinate difference is appreciable if one moves one's attention while fixating the little square but the difference is even more obvious if one moves fixation as well as attention.

The 22% and 28% patches look determinately different if one is attending to and foveating (looking right at) each in turn. So if representationism is true, there can be veridical representational contents of 22%-28% only if the phenomenal precision of the percepts of the patches seen attended and foveated is narrower than the phenomenal precision of at least one of

the percepts seen in the periphery with only one attended. This is a version of what I called the phenomenal precision principle in section 2. If two things look the same (veridically) when seen in peripheral vision with at least one unattended, but the same two things look determinately different—also veridically—when seen foveally and attentively, then the phenomenal precision of the attended and foveal percepts must be narrower than at least one of the prior percepts. And we can guess that it is the unattended prior percept that has to be less precise.

Recall that perceptual representations that are imprecise in that they attribute ranges can still be fine-grained. Suppose for example that a percept attributes a broad range of sizes to a gap of $.10^{\circ}$ – $.50^{\circ}$. That is a different representational content from $.11^{\circ}$ – $.51^{\circ}$, and that is different from $.12^{\circ}$ – $.52^{\circ}$. So our ability to see small differences can be based on absolute representation even if perception is imprecise. But if the representational contents of the foveal percepts almost totally overlap, as with $.11^{\circ}$ – $.51^{\circ}$ $.12^{\circ}$ – $.52^{\circ}$, how could those representational contents ground the determinately different phenomenologies?

Consider an analog for inattentive peripheral perception of color in which there is a red patch on one side and a blue patch on the other and the subject fixates in the middle. Suppose—and as far as I know this is science fiction—that there is some distribution of attention such that the two patches seen briefly and inattentively in the periphery can look the same and look red-blue and have the representational content red-blue. I don't mean reddish blue. I mean indeterminate as between central red and central blue or in between. They could be red, they could be blue, or they could be in between. Since attentive foveated percepts of red and blue in normal conditions are determinately different from one another (and from other colors) in phenomenology, the representational contents of red and blue seen foveated, attentively (and leisurely), would have to be more precise than the supposed contents seen peripherally, inattentively (and briefly). Otherwise there would be increasing precision in phenomenology without increasing precision in representational

content and representationism cannot allow that.

In short, representationism requires that inattentive peripheral perception be less precise representationally than attentive and foveal perception.

Now here is the striking fact: there is evidence that peripheral inattentive perception of many properties is not less representationally precise than foveal attentive perception. This conclusion conflicts with the application of the phenomenal precision principle to the cases at hand. I have already discussed the peripheral vs foveal aspect of this point and I will go over some of the evidence for the attentional component in section 11 below.

I will explain the argument just sketched in more detail. But first I must discuss a piece of the puzzle, the notion of a just noticeable difference.

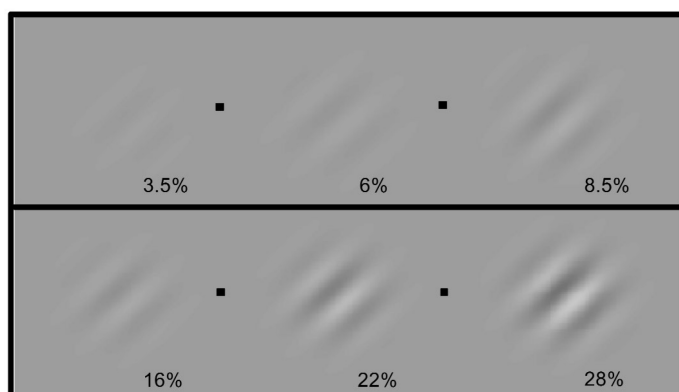


Figure 9: If one maintains fixation on one of the 4 little squares while varying attention to the patches on either side of the square, the patch to the left of the square seen with attention has the same appearance as the patch to the right without attention. I am grateful to Marisa Carrasco for this figure.

10 Just noticeable differences

A ubiquitous feature of perception is that perceptual discrimination is more fine-grained than perceptual identification. Even those with absolute pitch can identify perhaps 100 pitches (depending on exactly how absolute pitch is defined), but can discriminate many thousands of pitches from one another (Raffman 1995). Those who have absolute pitch are no better

than other musically literate people in pitch discrimination (Levitin 2005, 2008). Given the disparity between identification and discrimination one might wonder whether our ability to make fine grained perceptual discriminations misleads us as to the precision of our perceptual representations. It certainly seems to us that each of those thousands of pitches has a distinct phenomenology but maybe that judgment feeds more off of the phenomenology of discrimination of differences than off of the phenomenology of individual pitches.

I have argued that the phenomenology of perception does not allow for a large degree of imprecision.¹⁵ I appealed to the “just noticeable difference” of contrast of 2%. I said:

The representationist may retort that the point is not that the contents are fuzzy or represented indeterminately but that they are abstract relative to other contents, as determinables are to determinates, for example as red is to scarlet. But this line of thought runs into the following difficulty: the variation of 6% due to attention is way above the ‘just noticeable difference’ threshold, which for stimuli at these levels is approximately 2%. (Or so I am told. In any case, just looking at the stimuli in Figure 4 [Figure 9 here] shows that the difference is easily detectable. And you may recall that in the discussion of the tilt aftereffect, there was evidence that at higher levels of contrast, the increase due to attention was as much as 14%.) The point is that there is no single ‘look’ that something has if it is 22% plus or minus 6% in contrast. By analogy, consider the supposition that something looks as follows: rectangular or triangular or circular. That disjunctive predicate does not describe *one* way that something can look—at least not in normal perceptual circumstances (Block 2010, p. 52).

Jeremy Goodman (2013) has criticized my reasoning. He says:

¹⁵ In Block (2010). Actually, I spoke of “indeterminacy” and—mistakenly—of “vagueness” of perception.

Ned Block, when considering the hypothesis that perceptual appearances are ‘abstract relative to other contents, as determinables are to determinates, for example as red is to scarlet’, objects that ‘the variation of 6% due to attention is way above the “just noticeable difference” threshold, which for stimuli at these levels is approximately 2%’ (p. 35).

Goodman goes on to speak of “Block’s objection that our discrimination thresholds place an upper bound on the unspecificity of perceptual appearances...” (2013, p. 35). Although it may have sounded that way, I did not intend to claim that discrimination places an upper bound on either representational or phenomenal imprecision. But I think that a certain kind of discrimination is relevant to both imprecisions as I will explain.

An ability to discriminate between two observable magnitudes does not prove that one’s percepts of the magnitudes actually differ (in either representational content or phenomenology). One example that I have used to illustrate this point (Block 2007, p. 540) is the phenomenon of “beats” (alternating soft and loud sounds) caused by interference between guitar strings of very slightly different pitches even when the two pitches are phenomenally the same on their own. (The frequency of beats in response to two pure pitches is the difference in frequencies.) Another is the color border effects that allow one to see that two colors are different even when the colors themselves would not be distinguishable if separated slightly. Even for achromatic objects, slight differences are amplified by a well known border phenomenon that is responsible for the famous “Mach Bands” illusion. Goodman uses the example of two trees that look to have slightly different heights because of how far they stick up above the tree canopy. His point is that vision might represent overlapping but slightly different intervallic values, but one could also use the example to illustrate heights that don’t look at all different when seen separately while nonetheless allowing one to see a difference when seen next to one another.

It is intuitive to think that the way the visual system detects differences between one thing and another is by registering the properties of each thing separately and comparing those registrations. But this is not always the case: Differences are often detected via different processes than the processes that register the entities or properties that differ. Beats are produced by interference between two sound waves, allowing one to detect differences between sounds that would otherwise be inaudible.

As the examples just given illustrate, discrimination may be possible without any difference in the phenomenology of the individual percepts. Two pitches can be indiscriminable even if one knows they differ because of beats. However, there is no reason to think that specialized discrimination mechanisms are at work in the experiments described. Specialized discrimination mechanisms can be expected to depend on the specific features of the perceptual situations and so not robust to changes in the situation of the perception. For example, if you change your angle of view you might see the full vertical length of the trees but not their differential protrusion above the canopy. Border contrast effects are fragile—move the color samples just a bit apart and the effect vanishes. (This is nicely illustrated in the Wikipedia entry for “Mach Bands”.) However, the attentional effects I have been talking about apply to color, speed, size of a moving object, spatial frequency (stripe density), time of occurrence, flicker rate, motion coherence (the extent to which many moving items are going in the same direction), as well as to contrast and gap size. What is the chance that there is some specialized discrimination method at work for all these magnitudes? Most impressively, these effects can be exhibited in visual short term *memory*—that is, they don’t even require simultaneous perception. This was shown by Martin Rolfs & Marisa Carrasco using a different experimental paradigm than the ones so far discussed (2012; Rolfs et al. 2013). I won’t describe it except to say that the patches are compared in respect of contrast by comparing a patch seen earlier with a currently seen patch, and with similar results to those already described. As I mentioned in section 3,

a similar experiment shows that a perceived patch at one location in the visual field can be compared with a remembered patch at a different location with results that show the inhomogeneities in the visual field (Montaser-Kouhsari & Carrasco 2009). The likelihood that there is some method of comparison that does not depend on the individual percepts themselves but survives all these variations does seem slight.

So the kind of discrimination that is not based on specialized mechanisms of detecting differences independently of registering absolute value can be used to make a better case.

But even if we can make very fine grained discriminations and even if the percepts involved in the discriminations are distinct, it does not follow that the percepts are not highly imprecise—as mentioned earlier. Suppose for example, that perceptions of contrast are so imprecise as to cover nearly all the range of contrasts. Consider a representation of contrast of 4%-98%. Still, 4.1%-98.1% would be another equally imprecise content that is nonetheless distinct from the first one. And so more generally discriminability has little in the way of immediate consequences for imprecision.

The notion of a just noticeable difference is not very useful for my purposes. Discrimination can be finer than absolute registration as in the case of beats. And strong ability to discriminate is compatible with a high degree of imprecision. Further, the notion of a just noticeable difference combining as it does, perception with cognition, allows the possibility of a difference in conscious percepts that is not cognitively accessible.

11 Absolute representation

The phenomenal precision principle tells us that if the phenomenology of perception is grounded in its representational content, then peripheral unattended perception must be more imprecise than foveal attended perception. This result applies to contrast, size, spatial frequency and some other properties but not location. However, experimental results to be described in the next section suggest that contrast perception is as precise in foveal attended perception as in

peripheral unattended perception. But what this evidence does not tell us is how precise they both are, i.e., whether both are relatively precise or relatively imprecise.

I mentioned a study by Mazviita Chirimuuta & David Tolhurst (2005a) that is relevant to the issue of how precise absolute representations of contrast are in foveal attended perception. Chirimuuta and Tolhurst have a behavioral result that shows that performance in classifying contrasts falls off sharply after 4 contrasts. They have a neural model of contrast identification that suggests that the brain is capable of representing only 4-5 contrasts and that this limit is compatible with very fine-grained discriminations. Chirimuuta's view is that the response probabilities in the visual system for contrasts are very broad, with the tails of every distribution covering much of the span of possible contrasts. (That is, there is a non-zero probability across almost the whole range of contrasts.) Contrasts can only be identified when the response is near the peak of the probability distribution but two responses can be compared when responses are in the tails so long as the tails do not overlap much.

I'll start with the behavioral result. She presented subjects with a number of patches of up to 8 grades of contrast that were labeled "1" through "8" in each sequence of trials. Subjects looked at the contrasts and labels for as long as they liked and could have a refresher any time in the midst of the experiment if they liked. They had to hold the pairs of digits and contrasts in working memory and assign numbers to contrast stimuli. Then, patches were presented for half a second and subjects had to try to give the digit label. Performance was good up to 4 items and fell off drastically for larger sets.

Performance on 4 contrasts was near perfect. Then when new contrasts outside the original range were added, performance fell off, even for the original 4 contrasts. This is a pattern often seen in working memory experiments. For example, wild monkeys participated in an experiment in which an experimenter sets up two buckets and ostentatiously places, one at a time, a number of pieces of apple in each bucket. For example, there might be 4 in one

bucket and 3 in the other. The result is that for numbers of slices of 4 or less, monkeys reliably go to the bucket with more but with more than 4 items, performance falls off to chance (Barner et al. 2008; Hauser et al. 2000). Human infants show similar results with a limit closer to 3 (Feigenson et al. 2002).

The number 4 figures in working memory experiments in which subjects are asked to remember digits but are given another simultaneous distraction task to prevent overt strategies of “chunking” digits into units. Subjects can typically remember about 4 digits. In a completely different paradigm, George Sperling showed subjects a grid of letters briefly (1960). Subjects often said they could continue to see all or almost all the items faintly after the patch disappeared. (This kind of image has been called a visual “icon”.) When the grid had 3 rows of 4 items, and subjects were asked to recite as many letters as they could, they could name 3-4 letters. However Sperling gave subjects a cuing system: a high tone for the top row, a medium tone for the middle row and a low tone for the bottom row. When cued, subjects could report 3-4 from any given row.

In a different paradigm, honeybees were trained on a maze in which they had to choose to go either left or right at a T-junction to get a reward. At the entrance of the maze there were dots on each side and the bees had to choose the side with more dots to get the reward. The bees could learn to choose 4 rather than 3 but not 5 rather than 4 (Gross et al. 2009).

The working memory significance of roughly 4 items is so ubiquitous that it stimulated an article called “The magical number 4 in short-term memory: A reconsideration of mental storage capacity” (Cowan 2001). Up until 5-10 years ago, “slot” models of working memory were popular. I think it would now be agreed that roughly slot-like behavior emerges from an underlying working memory system in which there is a pool of resources that is distributed over items differently depending on number and complexity (Ma 2014). George Alvarez & Patrick Cavanagh (2004) suggested that there might be a limit of around 5 items of ideally simple structure but Alvarez’s recent work sug-

gests a more complex picture in which there are a variety of components of working memory that may independently fit a more slot-like or a more pool-like structure (Suchow et al. 2014). Slot-like working memory depends on simple stimuli that are hard to confuse with one another. Stimuli that have shown slot-like behavior include alphanumeric characters, horizontal/vertical rectangles and colors that differ substantially from one another. (I am indebted to conversations with Weiji Ma on this topic.)

So I would suggest that Chirimuuta’s behavioral result probably depends on the fact that subjects had to hold a number of pairs of digits and contrasts in mind in order to categorize the next contrast. (You could try it yourself for say 5 lengths.) They did well up to 4 such pairs and then performance declined radically. The article contains an anecdote that further supports this idea:

DJT [one of the subjects and experimenters] performed an experiment in which 4 contrasts of grating were chosen that were close together whilst still allowing near-perfect identification performance over 50 trials of each: 1, 8, 18 and 27 dB. [Note from NB: this is a different way of quantifying contrast than the percentages used here.] In the 50 trials of each contrast, 1 error of identification was made for each of the 8 and 18 dB gratings. Then, two more contrasts were added to the stimulus set at the lower end (40 and 50 dB); contrast 40 dB should have been easily discriminable from 27 dB. In fact, addition of contrasts 40 and 50 dB resulted in an increase in the errors of identification of the original set of four contrasts over 50 trials of each (8dB – 2 errors; 18dB – 9 errors; 27dB – 6 errors). (Chirimuuta & Tolhurst 2005a, p. 2965)

There are two notable aspects of this anecdote: first, performance over 50 trials of each of 4 contrasts were near perfect despite the fact that the gratings covered only part of the spectrum of contrasts. This suggests that the limit of 4 does not have to do with representations of con-

trast per se. The second aspect is that in this case as in so much of the work on working memory, adding more possibilities to a set of 4 decreases performance in the original set of 4. I conclude that the behavioral result probably has more to do with working memory than with any limit on perception.

Chirimuuta's second result, the one that motivates the idea that visual representations of contrast are so indeterminate that only 4-5 levels of identification are possible, is the modeling result based partly on data from monkey V1 neurons. (V1 is the first cortical area that processes vision, the lowest level of the visual system.) The striking fact about this result is that it does not concern working memory at all or indeed any kind of memory. It is only concerned with perceptual representation in V1. The model of V1 neurons comes from another paper that is concerned with the "dipper function", a notable curve shape in which one contrast stimulus is "masked"—diminished by the processing of another stimulus that follows right after it (Chirimuuta & Tolhurst 2005b). The model predicts that V1 can represent 4 contrasts perfectly with a sharp fall-off at 4, with a capacity to represent slightly more than 5 items.

However, the model based on V1 neurons gets some important facts wrong, for example it predicts poorer performance at high and low contrasts, whereas people actually do better at high and low contrasts. A version of the model with some postulated features that are not based on anything neural can get that right. However, this "curve fitting" approach deprives the model of the neurophysiological support that motivated the original model. Another problem with the model is that what is predicted is "mutual information" shared between contrast stimuli and V1 responses of 2.35 bits. Mutual information is a measure of shared information—in this case between stimuli and V1 neurons. A mutual information value of 2 bits would allow 2^2 (=4) contrast identifications; a mutual information value of 3 bits would allow 2^3 (=8) identifications. This shared information, as Chirimuuta notes (Chirimuuta & Tolhurst 2005a, p. 2968), is "essentially looking at per-

fect, 100% performance." For this reason, mutual information is not very useful as a psychophysical measure. And as Chirimuuta notes, its utility is limited for another reason: it is a compressive measure and so large increases in neural activity can be expected to make small differences in information. The issue of 100% performance is especially troublesome since in perceptual systems no performance can be perfect. In particular the convention for a "just noticeable difference" is distinguishability 75% of the time. So it is difficult to know how to compare the absolute identification level of 2.35 bits with a more visually sensible visual identification level.

Further, our experience seems to conflict with the idea that we have distinct visual representations of only 4-5 contrasts. A good reproduction of Figure 8 seems to reveal 6 phenomenologically different contrasts even though the figure covers only a third of the range of contrasts. And the Carrasco results apply to many different parameters, gap size, spatial frequency, etc. You might test it out if you happen to be near a brick wall. Look at the height of one brick, two bricks, three bricks and four bricks. If you are close enough so that those sizes look different from one another, ask yourself whether there are other sizes that look different from all four of those sizes. If Chirimuuta's result applies more widely, the answer is no. It has to be said though that that sense of distinctness could be due to discriminatory abilities.

Whatever the facts are about how precise foveal attentive perception is, the next section presents evidence that it is not more precise than inattentive peripheral perception.

12 Attention may not increase representational precision

I said that if 2 things look the same when seen in peripheral vision with at least one unattended, but the same two things look determinately different when seen foveally and attentively, then the phenomenal precision of the attended and foveal percepts must be greater than at least one of the prior percepts. (As I

mentioned, the assumption of veridicality is required to justify the imprecise representational contents of the peripheral percepts.)

It is common for philosophers to claim that attention increases “determinacy” of perception (Boone 2013; Nanay 2010; Stazicker 2011a, 2011b, 2013; but not Speaks 2010). The relevant kind of determinacy as I have been saying is precision. But it will be useful to distinguish precision from other forms of determinacy. Responses to attended stimuli are certainly less variable than responses to unattended stimuli. And attention increases acuity in the sense of spatial resolution, e.g., the ability to distinguish one dot from two dots. I will argue that spatial attention may not increase precision even if it reduces variability and acuity, and that further, in a rationally designed system spatial attention would not be expected to increase precision.

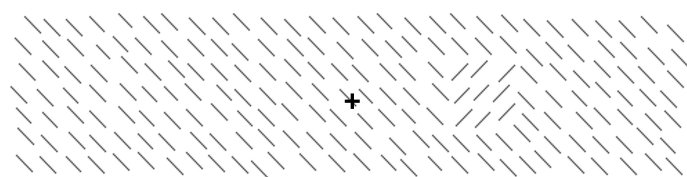


Figure 10: A textured figure used by Yeshurun & Carrasco (1998). Using stimuli like this one, stimuli were presented in which the square immediately to the right of the plus sign could appear at different eccentricities. When the resolution was low in peripheral areas, attention increased the subjects’ ability to detect the square. But when the resolution was high—nearer to the fixation point—attention *decreased* the subjects’ ability to detect the square because the increased resolution obscured the forest in favor of the trees.

Yeshurun & Carrasco (1998) showed that attention can increase resolution, making subjects (paradoxically) less likely to see the attended stimulus. For textured figures like the square to the right of the fixation plus sign in Figure 10, there is an optimal degree of resolution. If resolution is too high, the subjects miss the forest for the trees, failing to see the larger scale textured figures. Too low a resolution can cause subjects to miss the trees as well. Yeshurun and Carrasco presented textured figures at varying degrees of eccentricity. Since resolution is better for stimuli that are closer to the

fovea, this had the effect of presenting the figures at varying degrees of resolution. They also varied resolution by manipulating where subjects were attending, using cues of the sort described earlier. Putting together the contributions to resolution from eccentricity and attention, they were able to show that there were different optimal degrees of resolution for different figures.

One neural mechanism by which attention increases resolution is shrinking of the “receptive fields” of neurons in the visual system. Recall that a receptive field is the area of space that a neuron responds to. Resolution increases when neurons respond to smaller areas. Another mechanism is the shifting of receptive fields from adjacent areas that was mentioned earlier.

As I mentioned, the sensitivity of high “spatial frequency” channels increases—probably as a result of these mechanisms. Recall that spatial frequency in the case of a stripy stimulus like the “Gabor patches” used in many of the figures in this article (e.g., Figure 15) is a measure of how dense the stripes are. Boosting the sensitivity to high spatial frequencies makes resolution higher, thereby improving perception of textured figures when the resolution is too low and impairing perception when resolution is too high. The Yeshurun and Carrasco paper concerns exogenous attention. Later work (Barbot et al. 2012) shows that endogenous attention is more flexible, raising or lowering the sensitivities of high spatial frequency channels so as to improve perception.

I mention the increase in resolution and the sensitivity to high spatial frequencies in order to be sure that the reader is distinguishing these matters from an increase in precision.

Representational precision is a matter of how wide a range of values is allowed by the representational content, what values are compatible with the veridicality of the percept. (Phenomenal precision is a matter of “crispness” of the appearance.) One dot and two dots may look the same in peripheral vision even though we can clearly see the difference in foveal vision. That is a difference in acuity rather than a difference in precision. Increasing preci-

sion for representation is sharpening the representational content.

The relation between variability and precision is more complex. Imprecision is often cashed out in terms of reliable correlation between a representation and the world (Stazicker 2013). In that sense, since attention decreases variability it must increase precision. However, there are different sorts of noise. As we will see in the first experiment to be described below, attention may decrease noise across the whole spectrum without affecting what might be thought of as intrinsic variation in the signal and thus not increasing a kind of systematic precision. As I will explain, the experiment to be described helps us to precisify what precision comes to.

I will describe two experiments that will help to make the notion of precision more precise or at least concrete and will suggest that spatial attention does not narrow representational precision. Before I do that, let me say briefly why one should expect that spatial attention will not make the attended properties any more precise. Increasing precision normally involves suppression of responses outside the expected range. It would not make sense for a system to be designed to suppress some values without some indication of the irrelevancy of those values. Spatial attention tunes for spatial area, suppressing responses to other spatial areas (Montagna et al. 2009). So spatial attention can be expected to increase precision for spacial location but not for contrast, size, spatial frequency or speed.¹⁶ For feature-based attention, the opposite is true. If one is looking for the red thing, it makes sense to suppress sensitivity to other colors. Spatial attention should tune for space only and feature-based attention should tune for the property attended to.

The first experiment uses the “attentional blink”, a phenomenon in which there is a series of stimuli and two targets amid distractors. In part of the experiment, the targets were squares and the distractors circles. The general finding is that if the subject consciously sees the first

target square, and if the second target square is presented 200-400 ms after the first square, the subject will be much less likely to consciously see the second square. The mechanism has been shown to depend on the first target absorbing the subject’s attention so that there is insufficient attention to consciously see the second square. The second square is described as “blinked”, where the blinking deprives the square of attention. Asplund et al. (2014) used this technique with a paradigm in which the target squares were colored and in which subjects had to report the color of the second square by moving a mouse to click on a color wheel that had 180 colors on it. The idea is that the effect of attention on how intervalic the perceptual representation is could be assessed by examining the effect of the presence or absence of attention on the precision of subjects’ identifications of the color using the color wheel.

The experimental procedure is diagrammed in Figure 11. The subject saw a fixation point (lowest square on the left). Then there were 7-13 colored disks, then a target, T1, a square that was either black or white (RSVP = rapid serial visual presentation), then some number of colored disks, then another square, then 3 more disks. Then subjects reported the color of T2 using the color wheel. They got immediate feedback in how far off they were on identifying the color (in degrees on the color wheel) for 500 ms, then they indicated whether T1 was white or black. If the subject got T1 wrong, that trial’s report of T2 was disregarded. This design allowed for comparison of precision of reporting the color of T2 between trials in which attention was maximally reduced (T2 presented 2 items after T1, described as “lag 2”) with trials in which the lag was so long or so short that there was no attentional blink at all. The key result is that although the lag time was strongly correlated with the average correctness of the response (as always in the attentional blink), *the precision of the responses that were not random was not affected significantly*. The same experiment was done with faces using a slightly different form of the attentional blink. T1 was one of two faces that subjects had to recognize and the response wheel for T2 had a

¹⁶ This is oversimple since attention increases sensitivity to high spatial frequencies (Barbot et al. 2012).

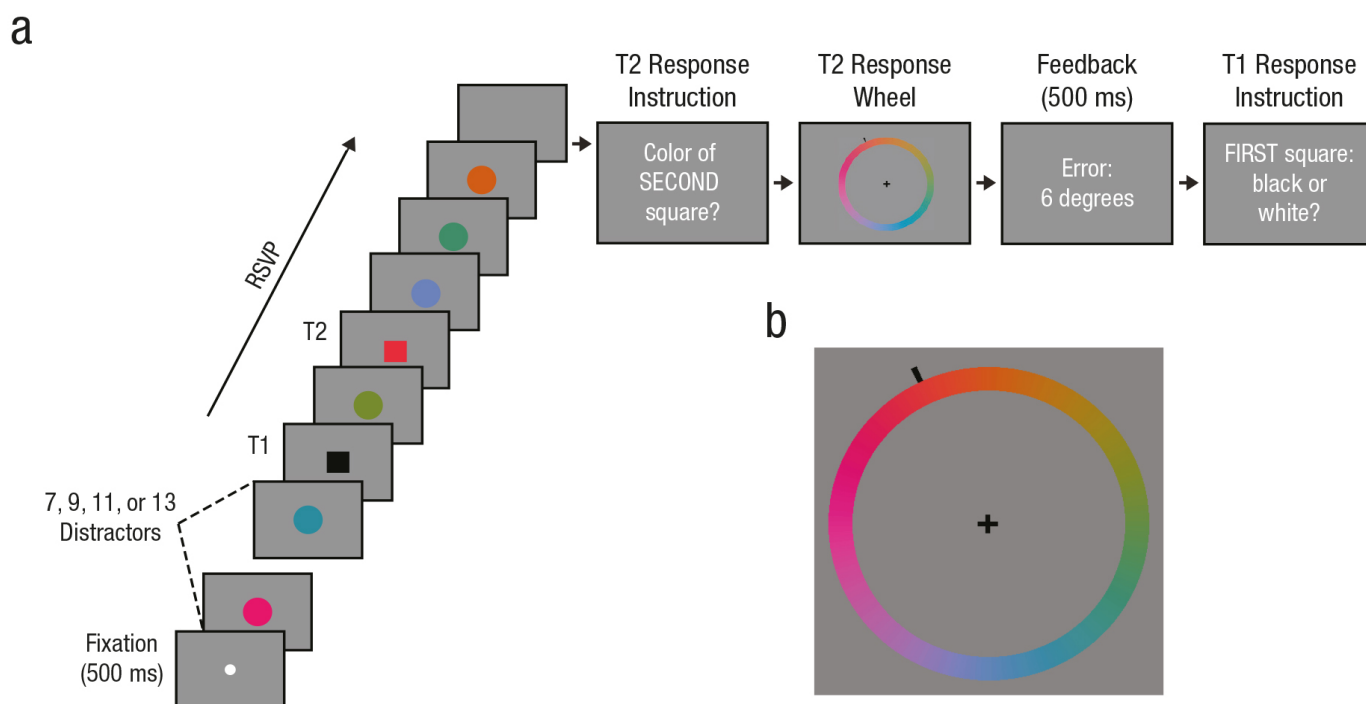


Figure 11: Procedure from [Asplund et al. \(2014\)](#). Understanding of this diagram is aided by color reproduction. Thanks to Chris Asplund for supplying this figure.

series of 150 face morphs based on 3 faces, with 49 intermediate faces interposed between them. The results were the same with faces as with colors. The key result for both studies is that the identification of T2 was either random (much more likely at the critical “lag 2” for an attentional blindness effect of 200-400 ms) or just as precise at lag 2 as at lag 8. Note that the experiment does not directly test the precision of any single percept. The assumption is that the precision of representation of a blinked color will be reflected in how tightly clustered the different responses are. [Asplund et al. \(2014\)](#) conclude (p. 6): “Across both stimulus classes (colors and faces) and experimental designs ..., we found that the reported precision of a target item is not affected in the AB [attentional blink], even though our paradigms had the sensitivity to detect such effects.”

But wait, you may ask: “Didn’t I say that attention decreases variability? And why is there supposed to be a difference between (the inverse of) variability and precision?” (Indeed, the inverse of “variance”, one measure of variability, is a common notion of precision.) The answer is that if you look at the raw data in

this experiment, the blinked color identifications are much more variable than the ones that are not blinked. However, the authors were able to show via modeling that the response distribution was a superimposition of two very different distributions. One distribution was uniform over the whole color wheel with no clustering around one color, whereas the second distribution was tightly clustered around the correct color, *just as tightly as when the color stimulus was not blinked*. They reasoned that the first (random) distribution represented cases in which the subject simply did not see the stimulus. However, when the subject did see the stimulus, the precision of the response was just as if it had not been blinked. (They considered and rejected a “variable precision” model that predicted the data less well; [van den Berg et al. 2012](#)). So overall variability of response is not a good guide to the precision of the representation. And this shows an important flaw in crude correlational approaches to precision. The precision of a perceptual representation should not be taken to be a matter of how well perceptual representation correlates with stimuli since what is really

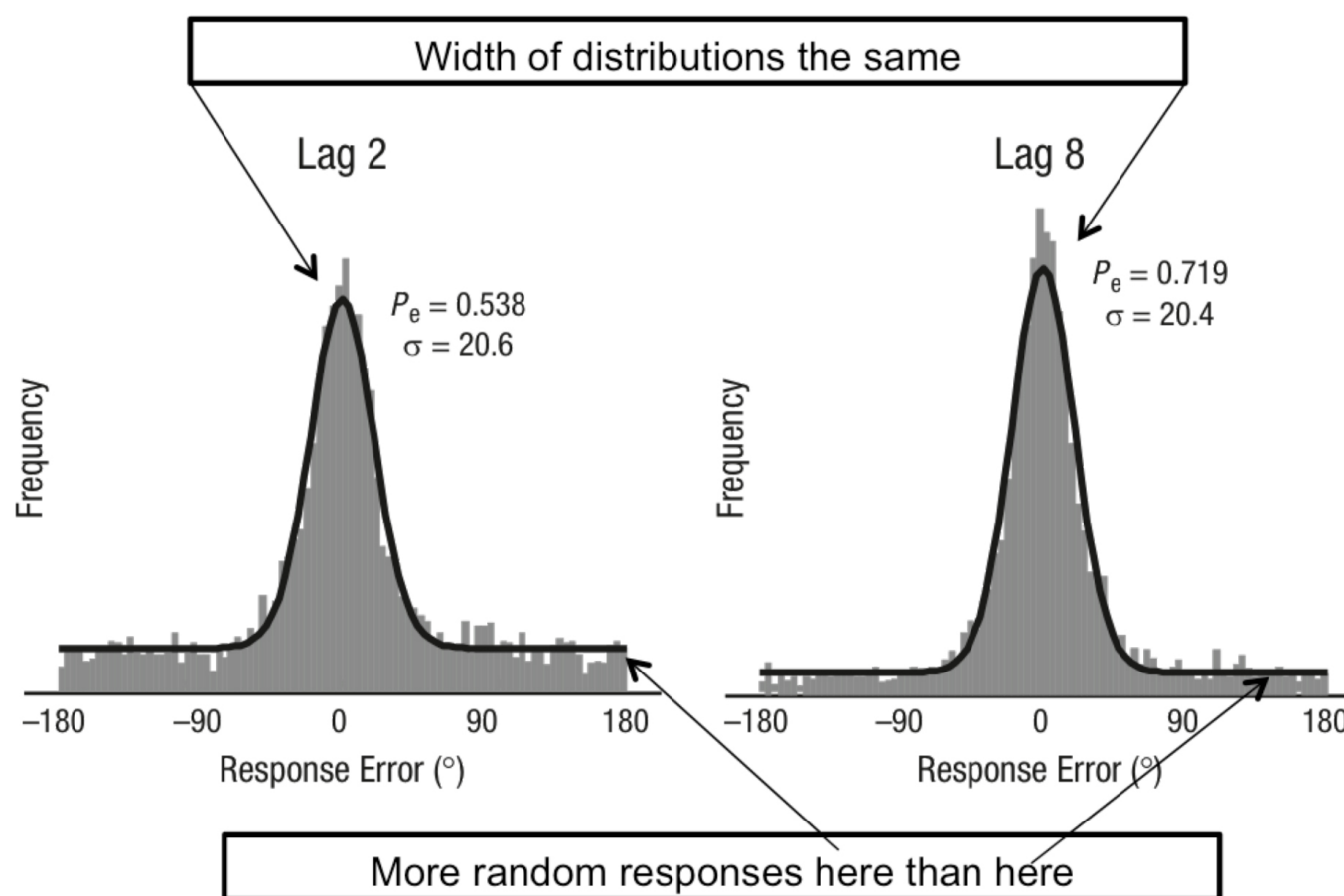


Figure 12: This is a modified form of a figure from [Asplund et al. \(2014\)](#). The figure compares response errors for lag 2—the value with the maximum effect of the attentional blink with lag 8—the value with the minimum attentional blink. What the figure shows is that the precision of the responses in which the subject actually saw the stimulus was the same. And the figure shows an increase in random responses for the blinked stimulus. Thanks to Chris Asplund for supplying the figure which has been modified here.

relevant is the cases in which the subject actually sees the stimulus.

This point is illustrated in Figure 12 in which the response error profile for lag 2 in which the attentional blink is most powerful is compared with the response error profile for lag 8 in which the attentional blink is least powerful. The widths of the distributions are the same. What differs is the number of random responses as indicated by the higher “tails” of the distributions.

Note the difference between precision and veridicality in this experiment. Precision is a matter of how tightly the responses cluster and veridicality is a matter of whether the responses cluster around the value of the item that was seen regardless of how tightly they cluster. If the color seen was focal red, responses could

pick out focal green in a very precise manner, but be non-veridical nonetheless. Conversely, the average of the responses might be the color seen (focal red) and thus the responses are on the average veridical even though the intervalic content is very wide.

An objector might say that the cases in which the blinked stimulus is reported in a non-random manner might be cases in which it was not in fact deprived of attention by the first percept. Imaging studies of the attentional blink do suggest a general deprivation of attention of the blinked stimulus ([Sergent et al. 2005](#)) but I don’t know of one that looks specifically at this issue. There are always potential confounds and the general remedy is to approach the same issue in more than one way. In the case of this result, the same con-

clusion has been reached by approaches that do not share vulnerabilities.

Another approach recorded from single neurons in a monkey visual area (V4) that is known to be sensitive to shape and form (David et al. 2008). Orientation tuning was not narrowed by spatial attention, but it was narrowed by attention to a specific orientation—feature-based attention. A recent review (Ling et al. 2014) summarizes this approach as follows:

Although initial physiological reports suggested that directing spatial attention to an item sharpens the band-width of orientation-selective cells in macaque visual area V4..., this was later shown not to necessarily be the case. Follow-up studies using a more sensitive measure for tuning band-width found no effect of spatial attention on the width of the orientation tuning function... Rather, these studies instead only found changes in the responsivity and baseline firing rate of neurons coding for the spatially attended location. Thus, the neurophysiological evidence appears to indicate that spatially attending to a location leaves a neuron's feature tuning unaffected.

A psychophysical study came to the same conclusion, that spatial attention boosts activation but not precision.

Ling et al. (2009) contrasted spatial and feature-based attention. The stimuli were random-dot cinematograms in which dots move in one direction or another for short distances. In the low noise condition shown on the left side of Figure 13, the dots show a high degree of coherence in that most of them move in the same direction. As noise increases, motion coherence decreases. Subjects had to make a series of judgments of the orientation of overall motion of these cinematograms. In the spatial attention version, they could be cued as to the place the stimulus would appear. In the feature-based attention version, they were cued to one of four directions of motion and had to report the observed motion as a clockwise or counterclockwise deviation from the cued motion. Ling et al.

were especially interested in comparing two different models for how attention boosts performance in detecting the direction of motion, using stimuli that could move in different directions. See Figure 14.

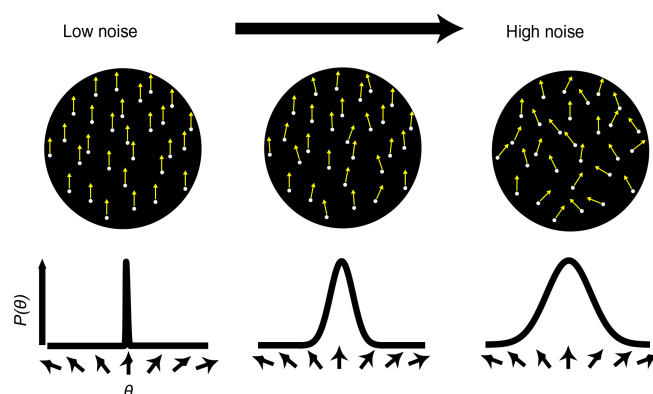


Figure 13: Random-dot cinematograms in which dots exhibit local motion. In the low noise condition, most of the dots are moving in the same direction. As noise increases, the spread of directions increases and motion coherence decreases. From an experiment comparing spatial attention with feature-based attention. With permission of *Vision Research*.

According to the gain model of (a), the response to the stimulus is increased as if the volume—i.e., the signal strength—were simply turned up equally for all movement-direction detectors. (Orientations of motion are indicated by arrows along the x-axis. The signal strength was turned up in the sense that the signal strength prior to attention is multiplied by a constant factor. For the values that are already high—i.e., at the peak—the multiplying a large value by a constant factor has a bigger effect than at the tails of the distribution where multiplying the constant factor times a zero yields zero.) According to the sharpening model of (b), the effect of attention is not to turn up the response but rather to suppress the irrelevant noise in the stimulus, narrowing the intervalic range of the response profile. These two models make different predictions for “threshold vs. noise” curves pictured in the bottom of Figure 14. The gain model predicts an increase in discriminability only when the external noise is low compared with internal noise. When external noise is low, there is a benefit to turning up the volume—even though the volume in-

creases both signal and external noise—since the effect of turning up the volume is to “swamp” the internal noise. (Internal noise is a blanket term for variation in the visual system, whatever makes visual responses vary even when the external signal remains the same.) As external signal and noise dominates the percept, internal noise decreases in importance. This kind of gain has a similar effect as decreasing the internal noise. If internal noise were zero, there would be no benefit at all in raising the volume on both the signal and the noise. The benefit of raising the volume however dwindles away as external noise increases since the increase in volume increases the effects of external noise too. This is indicated by the lowered threshold on the bottom left of (a) where the advantage in lowering the threshold decreases as external noise rises.

A different picture emerges from the model of the bottom right (b) where the benefit of tuning is greatest when external noise is greatest. (Note that if there is no external noise, tuning is of no benefit.) Thus the benefit should increase as noise increases as pictured in the bottom right (b). These models were tested by a procedure somewhat like that in Figure 5, except using voluntary attention. A line indicated where the subjects were supposed to attend and then a tone indicated that the stimulus was about to appear. Subjects could be cued to one of 4 locations where their task was to report the direction of motion of a stimulus. Sometimes there was a tone but no cue. The result was unequivocal: a pattern like that of the bottom left of Figure 14, indicating an effect of gain but no tuning. “The data showed that spatial attention yielded benefits strictly with low external noise, and no benefits with high external noise” (Ling et al. 2009, p. 1201). They also used the same setup with feature-based attention in which the subjects were cued with an indicator of what the direction of the stimulus would be. In this version, there was both tuning and gain, showing a hybrid of the patterns of a and b in Figure 14.

Again, spatial attention does not appear to narrow representational precision, contrary to the representationist position. This is graphic-

ally shown in the tuning model of Figure 14: suppression of values outside the selected value directly reduces precision. This is what does not happen with spatial attention.

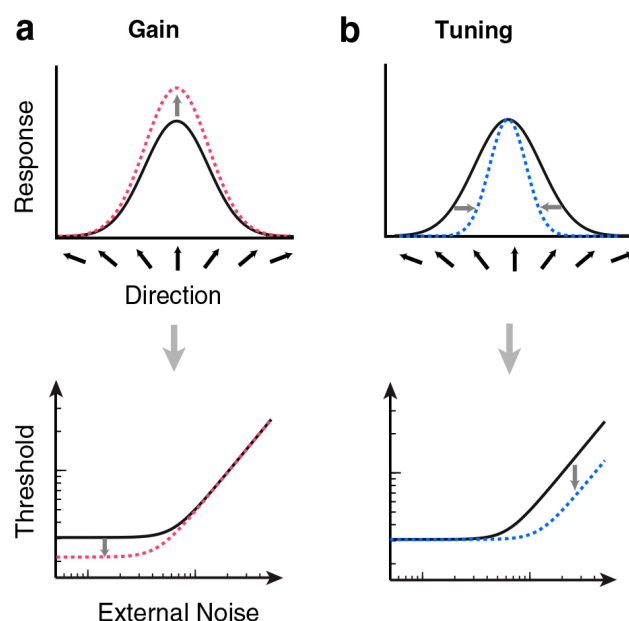


Figure 14: Two models of how attention boosts performance. According to the gain model indicated in the top left (a), the boost derives from increasing the firing of all directional feature detectors. The arrows along the x-axis indicate receptors for motion in different directions. The dotted lines represent the change due to attention (as compared with the solid lines). The tuning model at the top right (b) says performance is boosted by sharpening the response, decreasing the range of the intervalic content, as indicated by the narrowed shape of the dotted line. See the text for an explanation of the bottom diagrams. From Ling et al. (2009). With permission of *Vision Research*.

But why does this result concern representational precision rather than phenomenal precision? I considered an analog of this question concerning peripheral vision in section 2. There I noted that the anatomical asymmetries that are the probable basis of the inhomogeneities discussed are bound to affect unconscious perception in the same way as conscious perception. And a similar point applies here. The narrowing of receptive fields that is the main underlying mechanism of the attentional effects concerns perception *simpliciter* rather than conscious perception *per se*. As I mentioned earlier, spatial attention operates in unconscious per-

ception in a similar manner to conscious perception (Chica et al. 2011; Kentridge et al. 2008; Norman et al. 2013).

The dimensions used in both of the experiments described are “metathetic” as opposed to “prothetic” (Stevens & Galanter 1957). Prothetic dimensions have a zero point and intrinsic directionality, whereas metathetic dimensions have neither (Fuller & Carrasco 2006). Color saturation is prothetic because there is a zero point—achromaticity—and colors are more or less saturated. Hues are metathetic. At least for primary hues such as red and green, neither has more of any hue. Carrasco’s work shows that the attentional effects involved in increasing size, speed, flicker rate and the like work for prothetic dimensions like color saturation but not metathetic dimensions like hue (Fuller & Carrasco 2006). And that fact leads to the question of whether the conclusion that attention does not change precision depends on the magnitude tested being metathetic.

The studies on prothetic dimensions are not as easy to interpret as the ones I just described. One reason is that for metathetic dimensions, the psychological meaning of a difference is roughly the same throughout the dimension. A 90° shift in direction has roughly the same significance independently of the starting direction. But for prothetic magnitudes that is dramatically not so. A one inch change in a length of .01 inch has a different psychological significance than a one inch change in a length of one mile. Baldassi & Verghese (2005) give some evidence that spatial attention does not change the intervalic range of detection of contrast—a metathetic magnitude—though feature-based attention does narrow intervalic range.

The review I mentioned (Ling et al. 2014) surveys many different studies on this issue, concluding (references removed):

By and large, studies using psychophysical techniques to assess selectivity have converged on results that square quite nicely with the neurophysiological results...: feature-based attention to an item selectively changes psychophysical tuning curves..., while directing spatial attention to that

item leaves behavioral feature tuning untouched...

I mentioned that increasing precision normally involves suppression of responses outside the expected range. There is no reason for spatial attention to increase the precision of anything else other than spatial area. In particular, why would spatial attention suppress some directions of motion and not others? However if attention is directed towards motion in a certain direction (feature based attention) then increasing precision does make sense. The point applies equally to prothetic as to metathetic dimensions. Why should spatial attention tune for some values but not others of contrast or gap size given that tuning involves suppression of some range of contrasts or gap sizes. So there is good reason to expect these results to apply to prothetic dimensions.

Let me return to the issue of peripheral perception as compared with foveal perception. I mentioned the experiment that shows that discrimination of contrast in the periphery is as good as in the fovea. But there is an additional fact about peripheral vision, a phenomenon of “crowding” in which things lose the quality of “form...without losing crispness...” (Lettvin 1976). We can ignore crowding for the purposes discussed here so long as we confine ourselves to perception of what the visual system treats as single objects. For more on this, see Block (2012, 2013).

To conclude, there is evidence that attended and foveal perception can be greater in phenomenological precision without being greater in representational precision, contrary to representationism. In direct realist terms, there is evidence that attended and foveal perception can be greater in phenomenological precision without involving awareness of more precise environmental properties.

13 Abstraction and indeterminacy

The purpose of this section is to argue that the 6% difference between the attended 22% and unattended 28% patches underestimates the effect of attention. The reader who is will-

ing to take that on faith can skip to the conclusion.

I argued that since the attended 22% and unattended 28% patches look the same when seen in peripheral vision but look determinately different from one another when seen foveally and attentively, we can conclude that the precision of the phenomenal and therefore representational content of the attended foveal percepts must be greater than that of the prior percepts—if representationism is true and all the mentioned percepts are veridical. The reader may not be convinced however that the 22% and 28% patches do look determinately different when seen foveally and attentively. Perhaps the sense that they look different is a matter of an ability to discriminate rather than an appreciation of appearances that are determinately different.

The Carrasco lab experiments reported so far use stimuli that are 4° from the fixation point. But you might have noticed that when you fixated the bottom left square in Figure 9, you could also see the 28% patch to the far right. And some of the Carrasco lab's experiments have been done with 9° angle of separation.

If one combines the two different angles of separation as in Figure 15 an attended 16% patch looks the same in contrast as an unattended 28% patch, a larger difference than mentioned earlier for this absolute level of contrast. (The differences produced by attention increase with absolute level.) Of course the logic of the case is the same as before. I introduce it because I think it is easier to be sure that the patches in Figure 15 look determinately different when foveated and attended.

Of course there is a difference between the relations between the perceiver and the two patches—in the different angles of separation from the fixation point. Does that ruin the case for my purposes? Note that there was a difference in the relations between the perceiver and the two patches in the experiments of Figure 6 and Figure 7, namely one was on the left of fixation and the other was on the right. Why would there be a difference in relevance between left/right and number of degrees of peripherality?

Recall that imprecise contents were introduced in the first place via the following reasoning. An attended 22% patch looks the same as an unattended 28% patch. But both percepts with that same contrast phenomenology are veridical. In order for percepts with that phenomenology of contrast to be grounded in the representation of contrast, the imprecision of the representational content has to be at a minimum 22%-28% (inclusive of 22% and 28%). Suppose the representationist had said “No no, those phenomenologies are different since one is leftish and one is rightish so there can be no legitimate demand for a representational content in virtue of which they have the same phenomenology. That argument would look silly and be silly because we have an appreciation of how contrast looks independently of which side it is on. We can easily abstract the percept of contrast from a total percept of contrast on the left or contrast on the right. The sense of ‘abstract’ here is a question of appreciation of the phenomenology of contrast independently of perceived location: I speak of abstraction because location is abstracted away from.”

I suggest that the same reasoning applies to Figure 15 even though the difference in peripherality is causally implicated in producing the apparent contrast. The point is that we have an appreciation of that contrasty look independently of degree of peripherality and can appreciate that the two patches look the same in contrast when I am attending to the one on the left. The point is that a 16% patch can look the same in contrast as a 28% patch with the right distribution of attention and we need a representational account of what it is in virtue of which these apparent contrasts are the same. And with respect to that issue there is nothing illicit about comparing 4° with 9°.

The issue of abstraction I just mentioned comes up often in discussions of problem cases for representationism. Consider the phenomenal difference in seeing the round rim of a drinking glass and feeling it with one's hand. Both are percepts of one property, circularity, but the phenomenology is different. How can representationists cope with this case? Michael Tye (1995, p. 157; 2000, p. 93-95) has noted

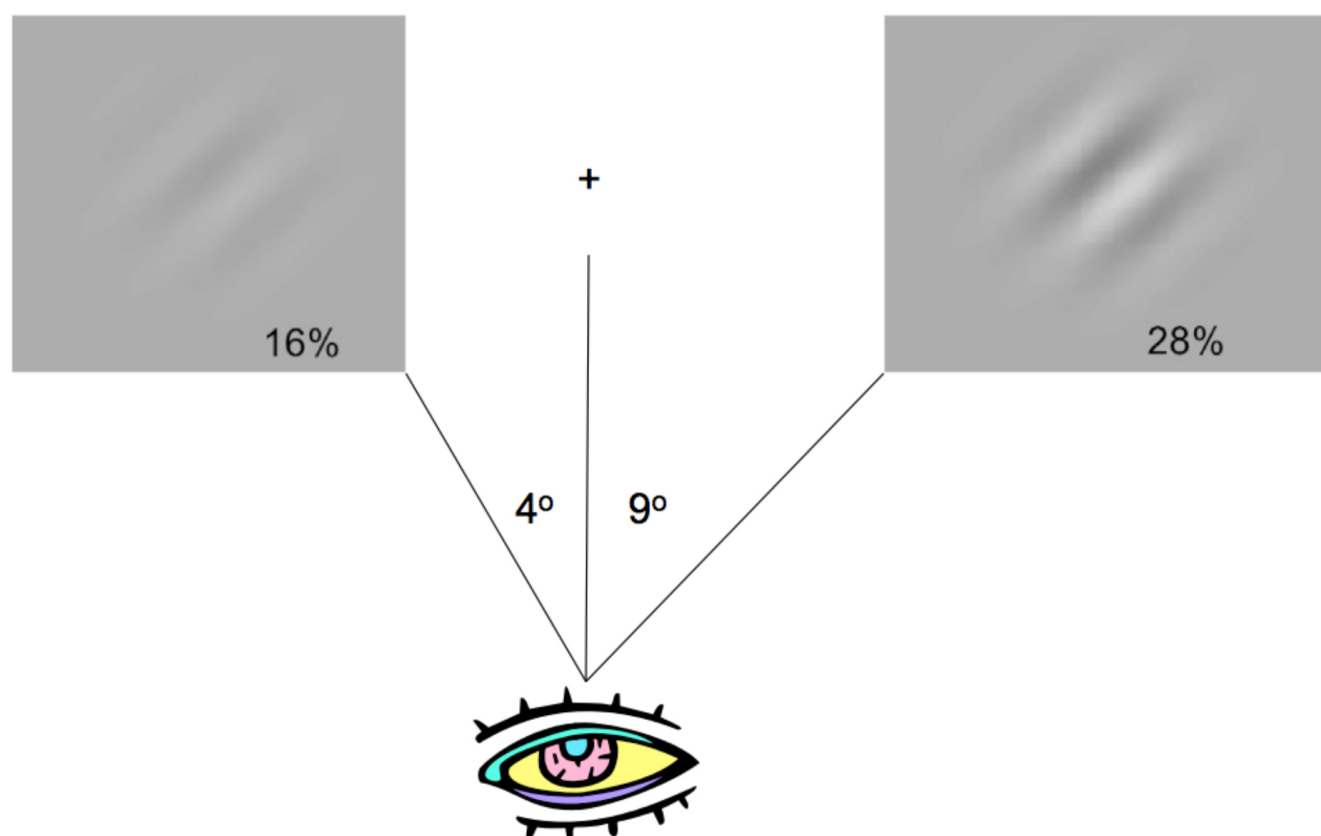


Figure 15: If you fixate at the “+” sign and attend to the left patch, it should look approximately equal in contrast to the right patch. My thanks to Jared Abrams for help in constructing this figure.

that the “total” percepts involve representation of different properties. These “collateral” properties might be shininess for the visual percept of the circularity and temperature for the tactile experience. The difference between the percepts can be blamed on the perceptions of these different properties. That is, what we are visually representing is circularity-&-shininess and what one is tactually representing is circularity-&-coldness. Can one abstract the visual impression of circularity from the total visual percept? Can one abstract the tactile impression of circularity from the total tactile percept? Tye says he cannot make sense of such abstraction. However, our ability to abstract shape from location on the right vs the left suggests the Principle of *Spatial* Abstraction: perceptual placing of a feature at a location can be abstracted from the perception of the location. I have a visual appreciation of the color of an object even as it moves, changing location. To the extent that this principle is accepted it licenses the use of Figure 15 in the

premise that the contrast percepts are determinately different.

14 Conclusion

I can now summarize the overall argument. First, the short version. The 22% patch and the 28% patch look different when foveated and attended one after the other. However, fixating in between them and attending to the 22% patch, they look the same. How can this be explained representationally without supposing that the precision of attentive foveal vision is narrower than that of inattentive peripheral vision? As before, this is a burden of proof argument that does not explicitly utilize the idea of phenomenal precision.

And as before, here is the long version:

1. The attended 22% patch and the unattended 28% patch, being the same in contrast-phenomenology are the same in contrast-representational contents.

2. Both are veridical.
3. The contrast attributed by vision to the two patches has a minimum span of 22%-28%.
4. Attended and foveal percepts of 22% and 28% (seen sequentially) are determinately different in phenomenology.
5. Phenomenal precision principle: the phenomenal precision of the percepts of the patches seen attended and foveally is narrower than the phenomenal precision of at least one of the percepts seen in the periphery with only one attended. And it is plausible to suppose it is the unattended percept that has the wider precision.
6. So the phenomenal precision of the attended foveal percepts must be narrower than at least one of the peripheral percepts (probably the unattended one).
7. Representationism requires that a difference in phenomenal precision be grounded in a commensurate difference in representational precision.
8. So representationism requires that the precision of the foveal attended percepts be narrower than at least one of peripheral percepts. We have already seen that peripherality *per se* probably does not decrease precision so if precision is decreased, it probably is due to withdrawal of attention. But empirical results suggest that withdrawal of attention does not decrease precision.
9. Conclusion: there is some reason to think that the phenomenology of perception is not grounded in its representational content.

Thus, for the perception of some properties, we have reason to believe that the representational content of perception neither grounds nor is grounded by the phenomenology of perception.

I argued that an attended .20° gap looks the same in respect of size as an unattended .23° gap. The comparative percept—the gaps looking the same—is illusory. But what about the percepts of each gap, considered separately? I argued that we would need a good reason to suppose that one but not the other is illusory and that the view that that both are illusory would undermine the notion of representational content altogether. I said that both are (or

rather can be in normal circumstances) veridical. A similar point applies to the version of the experiments involving contrast in which an attended 22% patch looks the same in contrast as an unattended 28% patch. If the two patches look the same and if looking the same is a matter of sameness in representational content, and if the percepts are veridical, the size properties the patches are represented as having must be intervallic. And the interval—an index of precision—must be wide enough to encompass both patches. So the representational content has to have a precision range of 6%. And further considerations I mentioned suggest a range of 12%. The phenomenal precision principle says if percepts of 22% and 28% are phenomenally the same with one unattended in peripheral vision but determinately different when attended and foveal, then the attended and foveal percepts must have a narrower phenomenal precision than at least one of the peripheral percepts. The 22% and 28% patches do look determinately different if foveated and attended. So the attended and foveal percepts must have a narrower phenomenal precision than one of the peripheral percepts. The only way that this can happen on the representationist point of view is if one of the peripheral representational content is less precise than the foveal attended content. But experimental results that I cited suggest that may not be true. It may not be true of foveal vs peripheral vision independently of attention, and it may not be true for attention independently of foveal vs peripheral perception.

In the section on inhomogeneities of the visual field, I mentioned a route to the same conclusion based on introspection. And I will update that point to include attention. The more introspective route is this: it is natural to feel that the phenomenology of seeing the contrast between the lines and spaces on a piece of lined paper attentively and foveally differs in precision from seeing the same lines inattentively and peripherally. The foveal attentive percept seems more “crisp” than the inattentive peripheral percept. As we have seen, location is indeed represented more precisely but the same is not true for other properties such as hue or

contrast. If this intuitive judgment is correct, there is introspective evidence for a discrepancy between the precision of phenomenology and the precision of representational content.

As I mentioned at the outset, the phenomenal precision principle needs more clarification and justification. It depends on notions of overlapping and of determinately different phenomenologies that are not as clear as one would like. My rationale is that if any advance in understanding of the phenomenology of perception is possible, it will have to start with underdeveloped ideas. I believe that there is enough in these ideas to give some credence to the conclusion. A second issue is whether the percepts that I say are determinately different in phenomenology really are.

The reader will have noticed that for the experimental results I have discussed it can often be difficult to figure out what aspects of the results concerned visual phenomenology and what aspects concern visual representation. As I mentioned earlier we have a real science of perception but very little science of the phenomenology of perception. If we are ever to turn what we know about perception into a scientific approach to the phenomenology of perception, we have no alternative but to start with some vague intuitive notions and proceed from there.

Although there are some loose ends, I think I have said enough to suggest a disconnect between the representational content of perception and what it is like to perceive.

Acknowledgements

I am grateful to Worth Boone, Tyler Burge, Marisa Carrasco, Jeremy Goodman, Eric Mandelbaum, John Morrison, Susanna Siegel, James Stazicker, Thomas Metzinger, Jennifer Windt and two anonymous reviewers for the Open MIND Project for comments on an earlier draft. And I am especially grateful to Jeremy Goodman and James Stazicker for discussion of these topics.

Glossary

Acuity	Also known as spatial resolution-- is the ability to resolve elements of stimuli. Common measures in the case of vision are the extent to which the subject can distinguish one dot from two dots, detect a gap between two figures, determine whether a rotating figure is rotating clockwise rather than counter-clockwise, ascertain whether two line segments are co-linear, distinguish a dotted from a solid line or detect which side of a Landolt Square a gap is on.
Attention	William James (1890, p. 404) famously said attention “...is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others.” Except for the exclusion of unconscious attention, most scientists would accept something like that characterization today. Spatial attention is attention directed to portion of environmental space and is distinct from attention to a thing or a property.
Content	See representational content.
Contrast	Contrast in an environmental layout is often defined as the average difference in luminance between light and dark areas. (Luminance is the amount of light reflected.) More specifically, it is the luminance difference between the lightest and darkest areas divided by the sum of those luminances. There are alternative ways of defining the notion but the differences won’t matter here.
Determinately different	For items to look determinately different in contrast, their contrast phenomenologies cannot be almost completely overlapping. I noted that this notion makes sense from a representationist perspective. I said that if one patch is represented as 10%-30% in contrast and another patch as 10.5%-30.5% the representationist would need a good reason to deny that the phenomenologies almost completely overlap. Given that representationism is committed to phenomenal precision and phenomenal overlap, it is legitimate to assume them in an argument against representationism.
Diaphanousness	G. E. Moore (1903) famously said “... the moment we try to fix our attention upon consciousness and to see what, distinctly, it is, it seems to vanish: it seems as if we had before us a mere emptiness. When we try to introspect the sensation of blue, all we can see is the blue; the other element is as if it were diaphanous ...”
Direct realism	The view that the phenomenal character of perceptual experience is grounded in direct awareness of objects and properties in the world.
Endogenous attention	Endogenous attention is voluntary—what people often mean by “paying attention”.
Exogenous attention	Exogenous spatial attention is attention that is attracted, automatically by a highly visible change. It is sometimes referred to as “transient” attention, whereas endogenous spatial attention is “sustained”. Exogenous spatial attention peaks by 120 ms after the cue, whereas endogenous spatial attention requires at least 300 ms to peak and has no known upper temporal limit.
Fixation	To fixate a thing or area of space is to point your eyes at it.
Fovea	The fovea is the high density center of the retina. Foveal vision is the only vision that can be 20/20. If you hold your hand at arm’s length, your foveal perception encompasses about double the width of your thumb.

Gabor patches	The fuzzy (actually sinusoidal) grids in Figure 1 and other figures.
Grounding	phenomenology is grounded in representational content just in case it is in virtue of the representational content of an experience that it has the phenomenology it has.
Identity formulation of representationism	What it is for an experience to have a certain phenomenal character is for it to have a certain representational content.
Landolt Square	See Figure 2.
Phenomenal precision principle	(one form) If two things look the same in peripheral vision but determinately different in foveal vision, then the phenomenal precision of foveal vision is narrower than that of peripheral vision.
Phenomenal precision	As with everything phenomenal, nothing like a definition is possible. The best you can do is use words to point to a phenomenon that the reader has to experience from the first person point of view. The experience of a color as red is less precise than the experience of a color as crimson. According to representationism, phenomenal precision is just the phenomenology of the precision of representational content. We experience a percept with representational content of 10%-20% as having more precision than we experience a percept with representational content 10%-30%. For a direct realist, phenomenal precision is just the precision of the properties we are directly aware of. We can be directly aware of properties with different precisions, for example, crimson, or alternatively red. Similarly we can be directly aware of a 10%-20% contrast property and also a 10%-30% contrast property and the difference constitutes a phenomenal precision difference.
Prothetic vs metathetic	Prothetic dimensions have a zero point and intrinsic directionality, whereas metathetic dimensions have neither.
Receptive field	In vision, the receptive field of a neuron is the area of space that a neuron responds to. In tactile perception the receptive field of a neuron is often gauged physiologically—the field of sensory receptors that feed to that neuron.
Representational content	Condition of veridicality. A simple percept consists of a representation of an environmental property and a singular element that picks out an individual item (Burge 2010). The representational content is satisfied when the referent of the singular element has the property represented by the property-representation.
Representational Precision	The precision of a representation is a matter of the intervalic range. For example, the precision of a representation of contrast of 10%-20% is narrower than a representation of 10%-30%. Precision in the sense used here is not a matter of indeterminacy of interval borders.
Spatial frequency	A measure of how closely spaced light and dark areas are. One could think of it with regard to the Gabor patches as a matter of stripe density.
Supervenience formulation of representationism	If phenomenology supervenes on representational content, there can be no difference in the phenomenology of perception without a difference in its representational content.
Veridicality	The veridicality of the most basic percept representations is a matter of the item referred to by the singular element having the property represented by the property representation.

References

- Abrams, J., Nizam, A. & Carrasco, M. (2012). Isoeccentric locations are not equivalent: The extent of the vertical meridian asymmetry. *Vision Research*, 52 (1), 70-78. [10.1016/j.visres.2011.10.016](https://doi.org/10.1016/j.visres.2011.10.016)
- Afraz, A., Pashkam, M. & Cavanagh, P. (2010). Spatial heterogeneity in the perception of face and form attributes. *Current Biology*, 20 (23), 2112-2116. [10.1016/j.cub.2010.11.017](https://doi.org/10.1016/j.cub.2010.11.017)
- Alvarez, G. & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science*, 15 (2), 106-111. [10.1111/j.0963-7214.2004.01502006.x](https://doi.org/10.1111/j.0963-7214.2004.01502006.x)
- Anton-Erxleben, K. & Carrasco, M. (2013). Attentional enhancement of spatial resolution: Linking behavioural and neurophysiological evidence. *Nature Reviews Neuroscience*, 14, 188-200. [10.1038/nrn3443](https://doi.org/10.1038/nrn3443)
- Anton-Erxleben, K., Henrich, C. & Treue, S. (2007). Attention changes perceived size of moving visual patterns. *Journal of Vision*, 7 (11), 1-9. [10.1167/7.11.5](https://doi.org/10.1167/7.11.5)
- Anton-Erxleben, K., Abrams, J. & Carrasco, M. (2010). Evaluating comparative and equality judgments in contrast perception: Attention alters appearance. *Journal of Vision*, 10 (11), 1-22. [10.1167/10.11.6](https://doi.org/10.1167/10.11.6)
- (2011). Equality judgments cannot distinguish between attention effects on appearance and criterion: A reply to Schneider. *Journal of Vision*, 11 (13), 1-8. [10.1167/11.13.8](https://doi.org/10.1167/11.13.8)
- Anton-Erxleben, K., Herrmann, K. & Carrasco, M. (2013). Independent effects of adaptation and attention on perceived speed. *Psychological Science*, 24 (2), 150-159. [10.1177/0956797612449178](https://doi.org/10.1177/0956797612449178)
- Aristotle (1955). On dreams. In W. D. Ross (Ed.) *Aristotle. Parva naturalia*. Oxford, UK: Clarendon Press.
- Asplund, C., Fougner, D., Zughini, S., Martin, J. W. & Marois, R. (2014). The attentional blink reveals the probabilistic nature of discrete conscious perception. *Psychological Science*, 25 (3), 824-831. [10.1177/0956797613513810](https://doi.org/10.1177/0956797613513810)
- Baldassi, S. & Verghese, P. (2005). Attention to locations and features: Different top-down modulation of detector weights. *Journal of Vision*, 5 (6), 556-570. [10.1167/5.6.7](https://doi.org/10.1167/5.6.7)
- Barbot, A., Landy, M. & Carrasco, M. (2012). Differential effects of exogenous and endogenous attention on 2nd-order texture contrast sensitivity. *Journal of Vision*, 12 (8), 1-15. [10.1167/5.6.7](https://doi.org/10.1167/5.6.7)
- Barner, D., Wood, J., Hauser, M. & Carey, S. (2008). Evidence for a non-linguistic distinction between singular and plural sets in rhesus monkeys. *Cognition*, 107 (2), 603-622. [10.1016/j.cognition.2007.11.010](https://doi.org/10.1016/j.cognition.2007.11.010)
- Bayne, T. (2014). Gist! Paper presented at NYU October 29, 2014.
- Block, N. (2007). Overflow, access and attention. *Behavioral and Brain Sciences*, 30 (5-6), 530-542. [10.1017/S0140525X07003044](https://doi.org/10.1017/S0140525X07003044)
- (2010). Attention and mental paint. *Philosophical Issues: A Supplement to Noûs*, 20 (1), 23-63. [10.1111/j.1533-6077.2010.00177.x](https://doi.org/10.1111/j.1533-6077.2010.00177.x)
- (2012). The grain of vision and the grain of attention. *Thought*, 1 (3), 170-184. [10.1002/tht3.28](https://doi.org/10.1002/tht3.28)
- (2013). Seeing and windows of integration. *Thought*, 2, 29-39. [10.1002/tht.62](https://doi.org/10.1002/tht.62)
- (2014a). The Canberra plan neglects ground. In T. Horgan, M. H. Sabates & D. Sosa (Eds.) *Qualia and mental causation in a physical world: Themes from the philosophy of Jaegwon Kim*. Cambridge, MA: Cambridge University Press.
- (2014b). Seeing-as in the light of vision science. *Philosophy and Phenomenological Research*, 89 (3), 560-572. [10.1111/phpr.12135](https://doi.org/10.1111/phpr.12135)
- Boone, W. (2013). Range content, attention and the precision of representation. *Presentation at the Society for Philosophy and Psychology*, June 15, 2013, Brown University
- Botta, F., Lupiáñez, J. & Chica, A. (2014). When endogenous spatial attention improves conscious perception: Effects of alerting and bottom-up activation. *Consciousness and Cognition*, 23, 63-73. [10.1016/j.concog.2013.12.003](https://doi.org/10.1016/j.concog.2013.12.003)
- Burge, T. (2005). Disjunctivism and perceptual psychology. *Philosophical Topics*, 33 (1), 1-78. [10.5840/philtopics20053311](https://doi.org/10.5840/philtopics20053311)
- (2010). *Origins of objectivity*. Oxford, UK: Oxford University Press.
- Cameron, E., Tai, J. & Carrasco, M. (2002). Covert attention affects the psychometric function of contrast sensitivity. *Vision Research*, 42 (8), 949-967. [10.1016/S0042-6989\(02\)00039-1](https://doi.org/10.1016/S0042-6989(02)00039-1)
- Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research*, 51 (13), 1484-1525. [10.1016/j.visres.2011.04.012](https://doi.org/10.1016/j.visres.2011.04.012)
- Carrasco, M., Talgar, C. & Cameron, E. (2001). Characterizing visual performance fields: Effects of transient covert attention, spatial frequency, eccentricity, task and set size. *Spatial Vision*, 15 (1), 61-75. [10.1163/15685680152692015](https://doi.org/10.1163/15685680152692015)

- Carrasco, M., Williams, P. E. & Yeshurun, Y. (2002). Covert attention increases spatial resolution with or without masks: Support for signal enhancement. *Journal of Vision*, 2 (6), 467-479. [10.1167/2.6.4](#)
- Carrasco, M., Ling, S. & Read, S. (2004). Attention alters appearance. *Nature Neuroscience*, 7 (3), 308-313. [10.1038/nm1194](#)
- Carrasco, M., Fuller, S. & Ling, S. (2008). Transient attention does increase perceived contrast of suprathreshold stimuli: A reply to Prinzmetal, Long and Leonhardt (2008). *Perception and Psychophysics*, 70 (7), 1151-1164. [10.3758/PP.70.7.1151](#)
- Chalmers, D. (2004). The representational character of experience. In B. Leiter (Ed.) *The future for philosophy* (pp. 153-181). Oxford, UK: Oxford University Press.
- (2006). Perception and the fall from Eden. In T. S. Gendler & J. Hawthorne (Eds.) *Perceptual Experience* (pp. 49-125). Oxford, UK: Oxford University Press.
- Chica, A., Lasaponara, S., Lupiáñez, J., Doricchi, F. & Bartolomeo, P. (2010). Exogenous attention can capture perceptual consciousness: ERP and behavioural evidence. *NeuroImage*, 51 (3), 1205-1212. [10.1016/j.neuroimage.2010.03.002](#)
- Chica, A., Lasaponara, S., Chanes, L., Valero-Cabre, A., Doricchi, F., Lupiáñez, J. & Bartolomeo, P. (2011). Spatial attention and conscious perception: The role of endogenous and exogenous orienting. *Attention, Perception & Psychophysics*, 73 (4), 1065-1081. [10.3758/s13414-010-0082-6](#)
- Chica, A., Paz-Alonso, P., Valero-Cabre, A. & Bartolomeo, P. (2013). Neural bases of the interactions between spatial attention and conscious perception. *Cerebral Cortex*, 23 (6), 1269-1279. [10.1093/cercor/bhs087](#)
- Chirimuuta, M. & Tolhurst, D. (2005a). Accuracy of identification of grating contrast by human observers: Bayesian models of V1 contrast processing show correspondence between discrimination and identification performance. *Vision Research*, 45 (23), 2960-2971. [10.1016/j.visres.2005.06.021](#)
- (2005b). Does a Bayesian model of V1 contrast coding offer a neurophysiological account of human contrast discrimination? *Vision Research*, 45 (23), 2943-2959. [10.1016/j.visres.2005.06.022](#)
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36 (3), 233-253. [10.1017/S0140525X12000477](#)
- Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24 (1), 87-185. [10.1017/S0140525X01373922](#)
- Craig, J. & Johnson, K. (2000). The two-point threshold: Not a measure of tactile spatial resolution. *Current Directions in Psychological Science*, 9 (1), 29-31. [10.1111/1467-8721.00054](#)
- Cutrone, E., Heeger, D. J. & Carrasco, M. (in press). Attention enhances contrast appearance via increased input baseline of neural responses. *Journal of Vision*.
- Datta, R. & DeYoe, E. (2009). I know where you are secretly attending! The topography of human visual attention revealed with fMRI. *Vision Research*, 49 (10), 1037-1044. [10.1167/10.7.9](#)
- David, S., Haydon, B., Mazer, J. & Gallant, J. (2008). Attention to stimulus features shifts spectral tuning of V4 neurons during natural vision. *Neuron*, 59 (3), 509-521. [10.1016/j.neuron.2008.07.001](#)
- Dehaene, S. (2014). *Consciousness and the brain: Deciphering how the brain codes our thoughts*. New York, NY: Viking.
- Deubel, H., Irwin, D. & Schneider, W. X. (1999). The subjective direction of gaze shifts long before the saccade. In W. Becker, H. Deubel & T. Mergner (Eds.) *Current oculomotor research: Physiological and psychological aspects* (pp. 65-70). New York, NY: Plenum.
- Emmanouil, T. & Magen, H. (2014). Neural evidence for sequential selection of object features. *Trends in Cognitive Sciences*, 18 (8), 390-391. [10.1016/j.tics.2014.04.007](#)
- Feigenson, L., Carey, S. & Hauser, M. (2002). The representations underlying infants' choice of more: Object files vs. analog magnitudes. *Psychological Science*, 13 (2), 150-156.
- Feynman, R. P. (1988). *QED: The strange theory of light and matter*. Princeton, NJ: Princeton University Press.
- Fine, K. (2012). Guide to ground. In F. Correia & B. Schnieder (Eds.) *Metaphysical grounding: Understanding the structure of reality* (pp. 37-80). Cambridge, UK: Cambridge University Press.
- Fuller, S. & Carrasco, M. (2006). Exogenous attention and color perception: Performance and appearance of saturation and hue. *Vision Research*, 46 (23), 4032-4047. [10.1016/j.visres.2006.07.014](#)
- Gobell, J. & Carrasco, M. (2005). Attention alters the appearance of spatial frequency and gap effect. *Psychological Science*, 16 (8), 644-651. [10.1111/j.1467-9280.2005.01588.x](#)

- Goodale, M. A. & Murphy, K. (1997). Action and perception in the visual periphery. In P. Their & H.-O. Karnath (Eds.) *Parietal lobe contributions to orientation in 3-D space* (pp. 447-461). New York, NY: Springer.
- Goodman, J. (2013). Inexact knowledge without improbable knowing. *Inquiry*, 56 (1), 30-53. [10.1080/0020174X.2013.775013](https://doi.org/10.1080/0020174X.2013.775013)
- Gross, H. J., Pahl, M., Si, A., Zhu, H., Tautz, J. & Zhang, S. (2009). Number-based visual generalization in the honeybee. *PloS ONE*, 4 (1), 1-9. [10.1371/journal.pone.0004263](https://doi.org/10.1371/journal.pone.0004263)
- Hauser, M., Carey, S. & Hauser, L. (2000). Spontaneous number representation in semi-free ranging rhesus monkeys. *Proceedings of the Royal Society of London: Biological Sciences*, 267 (1445), 829-833. [10.1098/rspb.2000.1078](https://doi.org/10.1098/rspb.2000.1078)
- Herrmann, K., Montaser-Kouhsari, L., Carrasco, M. & Heeger, D. J. (2010). When size matters: Attention affects performance by contrast or response gain. *Nature Neuroscience*, 13 (12), 1554-1559. [10.1038/nn.2669](https://doi.org/10.1038/nn.2669)
- Hess, R. & Field, D. (1993). Is the increased spatial uncertainty in the normal periphery due to spatial undersampling or uncalibrated disparity? *Vision Research*, 33 (18), 2663-2670. [10.1016/0042-6989\(93\)90226-M](https://doi.org/10.1016/0042-6989(93)90226-M)
- Hill, C. (2009). *Consciousness*. Cambridge, UK: Cambridge University Press.
- Hohwy, J. (2013). *The predictive mind*. Oxford, UK: Oxford University Press.
- Hohwy, J., Roepstorff, A. & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, 108 (3), 687-701. [10.1016/j.cognition.2008.05.010](https://doi.org/10.1016/j.cognition.2008.05.010)
- Horgan, T. & Tienson, J. (2002). *Philosophy of mind: Classical and contemporary readings*. Oxford, UK: Oxford University Press.
- Itti, L. & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40, 1489-1506. [10.1.1.40.4697](https://doi.org/10.1.1.40.4697)
- James, W. (1890). *Principles of psychology*. New York, NY: Henry Holt.
- Kentridge, R., Nijboer, T. & Heywood, C. (2008). Attended but unseen: Visual attention is not sufficient for visual awareness. *Neuropsychologia*, 46 (3), 864-869. [10.1016/j.neuropsychologia.2007.11.036](https://doi.org/10.1016/j.neuropsychologia.2007.11.036)
- Kerzel, D., Zarian, L., Gauch, A. & Buetti, S. (2010). Large effects of peripheral cues on appearance correlate with low precision. *Journal of Vision*, 10 (11), 1-14. [10.1167/10.11.26](https://doi.org/10.1167/10.11.26)
- Kriegel, U. (2011). *The sources of intentionality*. Oxford, UK: Oxford University Press.
- (2013). The phenomenal intentionality research program. In U. Kriegel (Ed.) *Phenomenal intentionality* (pp. 1-26). Oxford, UK: Oxford University Press.
- Lettvin, J. Y. (1976). On seeing sidelong. *The Sciences*, 16 (4), 10-20. [10.1002/j.2326-1951.1976.tb01231.x](https://doi.org/10.1002/j.2326-1951.1976.tb01231.x)
- Levi, D. & Klein, D. (1996). Limitations on position coding imposed by undersampling and univariance. *Vision Research*, 36 (14), 2111-2120. [10.1016/0042-6989\(95\)00264-2](https://doi.org/10.1016/0042-6989(95)00264-2)
- Levitin, D. (2005). Absolute pitch: Perception, coding, and controversies. *Trends in Cognitive Sciences*, 9 (1), 26-33. [10.1016/j.tics.2004.11.007](https://doi.org/10.1016/j.tics.2004.11.007)
- (2008). Absolute pitch: Both a curse and a blessing. In M. Klockars & M. Peltomaa (Eds.) *Music meets medicine. Proceedings of the Signe and Ane Gyllenberg Foundation* (pp. 124-132). Helsinki, Finland: Signe and Ane Gyllenberg Foundation.
- Ling, S. & Carrasco, M. (2006). When sustained attention impairs perception. *Nature Neuroscience*, 9 (10), 1243-1245. [10.1038/nn1761](https://doi.org/10.1038/nn1761)
- Ling, S., Liu, T. & Carrasco, M. (2009). How spatial and feature-based attention affect the gain and tuning of population responses. *Vision Research*, 49 (10), 1194-1204. [10.1016/j.visres.2008.05.025](https://doi.org/10.1016/j.visres.2008.05.025)
- Ling, S., Jehee, J. & Pestilli, F. (2014). A review of the mechanisms by which attentional feedback shapes visual selectivity. *Brain Structure and Function*. [10.1007/s00429-014-0818-5](https://doi.org/10.1007/s00429-014-0818-5)
- Ma, W. J., Husain, M. & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, 17 (3), 347-356. [10.1038/nn.3655](https://doi.org/10.1038/nn.3655)
- Metzinger, T. (2003). *Being no one*. Cambridge, MA: MIT Press.
- Montagna, B., Pestilli, F. & Carrasco, M. (2009). Attention trades off spatial acuity. *Vision Research*, 49 (7), 735-745.
- Montaser-Kouhsari, L. & Carrasco, M. (2009). Perceptual asymmetries are preserved in short-term memory tasks. *Attention, Perception & Psychophysics*, 71 (8), 1782-1792. [10.3758/APP.71.8.1782](https://doi.org/10.3758/APP.71.8.1782)
- Moore, G. E. (1903). The refutation of idealism. *Mind*, 12 (48), 433-453.
- Morrison, J. (2013). Anti-atomism about color representation. *Noûs*. [10.1111/nous.12018](https://doi.org/10.1111/nous.12018)
- Nanay, B. (2010). Attention and perceptual content. *Analysis*, 70 (2), 263-269. [10.1093/analys/anp165](https://doi.org/10.1093/analys/anp165)
- Norman, L. J., Heywood, C. & Kentridge, R. (2013). Object-based attention without awareness. *Psychological*

- Science*, 24 (6), 836-843. [10.1177/0956797612461449](https://doi.org/10.1177/0956797612461449)
- Pautz, A. (2010). Why explain visual experience in terms of content? In B. Nanay (Ed.) *Perceiving the world: New essays on perception*. New York, NY: Oxford University Press. [10.1111/j.1520-8583.2007.00134.x](https://doi.org/10.1111/j.1520-8583.2007.00134.x)
- Perkins, R. & Bayne, T. (2013). Representationalism and the problem of vagueness. *Philosophical Studies*, 162 (1), 71-86. [10.1007/s11098-012-9990-8](https://doi.org/10.1007/s11098-012-9990-8)
- Phillips, I. B. (2011). Perception and iconic memory: What Sperling doesn't show. *Mind & Language*, 26 (4), 381-411. [10.1111/j.1468-0017.2011.01422.x](https://doi.org/10.1111/j.1468-0017.2011.01422.x)
- Prinz, J. J. (2012). *The conscious brain*. New York, NY: Oxford University Press.
- Raffman, D. (1995). On the persistence of phenomenology. In T. Metzinger (Ed.) *Conscious experience* (pp. 293-308). Paderborn, Germany: Ferdinand Schöningh.
- Ramachandran, V. S. & Hirstein, W. (1998). The perception of phantom limbs. *Brain*, 121 (Pt 9), 1603-1630. [10.1093/brain/121.9.1603](https://doi.org/10.1093/brain/121.9.1603)
- Reynolds, J. H. & Chelazzi, L. (2004). Attentional modulation of visual processing. *Annual Review of Neuroscience*, 27, 611-647. [10.1146/annurev.neuro.26.041002.131039](https://doi.org/10.1146/annurev.neuro.26.041002.131039)
- Reynolds, J. H., Pasternak, T. & Desimone, R. (2000). Attention increases sensitivity of V4 neurons. *Neuron*, 26 (3), 703-714. [10.1016/S0896-6273\(00\)81206-4](https://doi.org/10.1016/S0896-6273(00)81206-4)
- Reynolds, J. H. & Heeger, D. J. (2009). The normalization model of attention. *Neuron*, 61 (2), 168-185. [10.1016/j.neuron.2009.01.002](https://doi.org/10.1016/j.neuron.2009.01.002)
- Rolfs, M. & Carrasco, M. (2012). Rapid simultaneous enhancement of visual sensitivity and perceived contrast during saccade preparation. *The Journal of Neuroscience*, 32 (40), 13744-13752. [10.1523/JNEUROSCI.2676-12.2012](https://doi.org/10.1523/JNEUROSCI.2676-12.2012)
- Rolfs, M., Lawrence, B. & Carrasco, M. (2013). Reach preparation enhances visual performance and appearance. *Philosophical Transactions of the Royal Society B*, 368 (1628), 1-8. [10.1098/rstb.2013.0057](https://doi.org/10.1098/rstb.2013.0057)
- Schneider, K. A. (2006). Does attention alter appearance? *Perception and Psychophysics*, 68 (5), 800-814. [10.3758/BF03193703](https://doi.org/10.3758/BF03193703)
- (2011). Attention alters decision criteria but not appearance: A reanalysis of Anton-Erxleben, Abrams and Carrasco (2010). *Journal of Vision*, 11 (13), 1-8. [10.1167/11.13.7](https://doi.org/10.1167/11.13.7)
- Schneider, K. A. & Komlos, M. (2008). Attention biases decisions but does not alter appearance. *Journal of Vision*, 8 (15), 1-10. [10.1167/8.15.3](https://doi.org/10.1167/8.15.3)
- Schoenfeld, M. A., Hopf, J.-M., Merkel, C., Heinze, H.-J. & Hillyard, S. A. (2014). Object-based attention involves the sequential activation of feature-specific cortical modules. *Nature Neuroscience*, 17 (4), 619-624. [10.1038/nn.3656](https://doi.org/10.1038/nn.3656)
- Scholl, B. J., Noles, N. S., Pasheva, V. & Sussman, R. (2003). Talking on a cellular telephone dramatically increases 'sustained inattention blindness'. *Journal of Vision*, 3 (9), 156-156. [10.1167/3.9.156](https://doi.org/10.1167/3.9.156)
- Sergent, C., Baillet, S. & Dehaene, S. (2005). Timing of the brain events underlying access to consciousness during the attentional blink. *Nature Neuroscience*, 8 (10), 1391-1400. [10.1038/nn1549](https://doi.org/10.1038/nn1549)
- Shoemaker, S. (2007). *Physical realization*. Oxford, UK: Oxford University Press.
- Siegel, S. (2010). *The contents of visual experience*. Oxford, UK: Oxford University Press.
- (2013). Are there edenic grounds of perceptual intentionality? *Analysis Reviews*, 73 (2), 329-344. [10.1093/analys/ans150](https://doi.org/10.1093/analys/ans150)
- Sorenson, R. (2013). Vagueness. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. <http://plato.stanford.edu/archives/win2013/entries/vagueness>.
- Speaks, J. (2010). Attention and intentionalism. *Philosophical Quarterly*, 60 (239), 325-342. [10.1111/j.1467-9213.2009.617.x](https://doi.org/10.1111/j.1467-9213.2009.617.x)
- Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs*, 74 (11), 1-29. [10.1037/h0093759](https://doi.org/10.1037/h0093759)
- Stazicker, J. (2011a). Attention, visual consciousness and indeterminacy. *Mind & Language*, 26 (2), 156-184. [10.1111/j.1468-0017.2011.01414.x](https://doi.org/10.1111/j.1468-0017.2011.01414.x)
- (2011b). Attention, visual knowledge & psychophysics: Discriminating the determinable. *NYU Philosophy of Mind Discussion Group*.
- (2013). Attending, knowing and detecting signals. In J. Taylor (Ed.) *Consciousness and attention workshop, as part of 'Philosophy and psychology: Integrating research across domains'*. Durham, UK: Durham University Press.
- Stevens, S. S. & Galanter, E. H. (1957). Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology*, 54 (6), 377-411. [10.1037/h0043680](https://doi.org/10.1037/h0043680)
- Stoljar, D. (2004). The argument from diaphanousness. In M. Escudria, R. J. Stainton & C. D. Viger (Eds.) *Language, mind and world: Special issue of the Canadian Journal of Philosophy (Vol. 30)* (pp. 341-390). Edmonton, Canada: University of Alberta Press.
- Strasburger, H., Rentschler, I. & Jüttner, M. (2011). Peripheral vision and pattern recognition: A review. *Journal of Vision*, 11 (5), 1-82. [10.1167/11.5.13](https://doi.org/10.1167/11.5.13)

- Suchow, J., Fougnie, D., Brady, T. F. & Alvarez, G. (2014). Terms of the debate on the format and structure of visual memory. *Attention, Perception & Psychophysics*, 76 (7), 2071-2079. [10.3758/s13414-014-0690-7](#)
- Tong, J., Mao, O. & Goldreich, D. (2013). Two-point orientation discrimination versus the traditional two-point test for tactile spatial acuity assessment. *Frontiers in Human Neuroscience*, 7 (579), 1-11. [10.3389/fnhum.2013.00579](#)
- Treue, S. (2004). Perceptual enhancement of contrast by attention. *Trends in Cognitive Sciences*, 8 (10), 435-437. [10.1016/j.tics.2004.08.001](#)
- Turatto, M., Vescovi, M. & Valsecchi, M. (2007). Attention makes moving objects be perceived to move faster. *Vision Research*, 47 (2), 166-178. [10.1016/j.visres.2006.10.002](#)
- Tye, M. (1995). *Ten problems of consciousness*. Cambridge, MA: MIT Press.
- (2000). *Consciousness, color, and content*. Cambridge, MA: MIT Press.
- (2009). *Consciousness revisited*. Cambridge, MA: MIT Press.
- Valsecchi, M., Vescovi, M. & Turatto, M. (2010). Are the effects of attention on speed judgments genuinely perceptual? *Attention, Perception & Psychophysics*, 72 (3), 637-650. [10.3758/APP.72.3.637](#)
- van den Berg, R., Shin, H., Chou, W., George, R. & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences, USA*, 109 (22), 8780-8785. [10.1073/pnas.1117465109](#)
- Watzl, S. (forthcoming). Can intentionalism explain how attention affects appearances? In A. Pautz & D. Stoljar (Eds.) *Themes from Block*. Cambridge, MA: MIT Press.
- Wilkinson, S. (2014). Book review: The predictive mind. *Analysis*
- Wu, W. (2014). *Attention*. New York, NY: Routledge.
- Yeshurun, Y. & Carrasco, M. (1998). Attention improves or impairs visual performance by enhancing spatial resolution. *Nature*, 396 (6706), 72-75. [10.1038/23936](#)
- (1999). Spatial attention improves performance in spatial resolution tasks. *Vision Research*, 39 (2), 293-306. [10.1016/S0042-6989\(98\)00114-X](#)
- Zhang, S., Xu, M., Kamigaki, T., Do, J., Chang, W.-C., Jenvay, S., Miyamichi, K., Luo, L. & Dan, Y. (2014). Long-range and local circuits for top-down modulation of visual cortex processing. *Science*, 345 (6197), 660-665. [10.1126/science.1254126](#)

Phenomenal Precision and Some Possible Pitfalls

A Commentary on Ned Block

Sascha Benjamin Fink

Ground Representationism is the position that for each phenomenal feature there is a representational feature that accounts for it. Against this thesis, Ned Block has provided an intricate argument that rests on the notion of “phenomenal precision”: the phenomenal precision of a percept may change at a different rate from its representational counterpart. If so, there is then no representational feature that accounts for a specific change of this phenomenal feature. Therefore, Ground Representationism cannot be generally true.

Although the notion of phenomenal precision is intuitive, it is admittedly in need of clarification. Here I reconstruct Block’s argument by suggesting a way of estimating phenomenal precision that is based on the assumption that parts of perceptual wholes can share phenomenal features independently of their place in the whole. Understood like this, the overall argument shows what it is supposed to show.

A more thorough look at the notion of phenomenal precision suggests tension with Block’s other work: in order to be non-trivial, we have to accept that some of our phenomenality is not concrete, but only generic. Such “solely generic phenomenology”, however, is a position mainly held by opponents to Block’s *Access- vs. Phenomenal Consciousness*-distinction. Interpreting phenomenal imprecision as constituted by introspective imprecision does not suffice as a way out. It seems that phenomenal precision is either trivial, self-contradictory, or incompatible with Block’s position elsewhere. So some additional elucidation on this crucial notion is needed.

Keywords

Access consciousness | Grounding | Perception | Perceptual experiences | Phenomenal consciousness | Phenomenal unity | Phenomenality | Precision | Psychophysics | Representation | Representationism | Supervenience | Veridicality | Vision science

1 Introduction: Running representationism into the ground

Imagine yourself in an elevator. You press the button for the upmost floor when, all of a sudden, you smell something nauseating: a foul metallic odor permeates your nostrils and raises disgust until all your attention is focused (unfortunately) on this olfactory catastrophe. *How* it smells is not the question. The odor has a very determinate character—and it is funky! But *what* is it that you smell, what is this sensation *about*? Maybe you left a cheese sandwich

in your pocket and forgot about it? Maybe some wiring went faulty? Or the breaks? Maybe your colleague cut one out? Even though you don’t know *what* it is you are experiencing—what your experience is about, its content, or *representational* aspect—, you do know *how it is like* to smell this stench—you know its appearance, its character, its configurational¹ or

¹ For the distinction between representational and configurational aspects see Wollheim (1987). Nanay (2005) used these termini vis-à-vis

Commentator

Sascha Benjamin Fink

sfink@ovgu.de

Otto-von-Guericke-Universität
Magdeburg, Germany

Target Author

Ned Block

ned.block@nyu.edu

New York University
New York, NY, U.S.A.

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

phenomenal aspects. What is the relation between content and character in such percepts?

Representationists give the following answer to this question: all phenomenal features of an experience (its appearance or character) are dependent on its representational features (its content): *how* you experience is determined by *what* you experience. Ned Block (forthcoming 2015)² has provided a useful way of taxonomizing Representationists further: Identity-Representationism (IR) is the claim that the character of an experience is nothing but its content. However, content here cannot be more basic than character, because identity is symmetrical. Character then determines content just as content determines character, because both are just one and the same. Also, if we want to *explain* or *reduce* character to content, then IR is not the way to go, because reduction—unlike identity—is *asymmetric*, and so is explanation.³

That an experience's content is more basic and determines its character (but not *vice versa*) can be captured in two ways: Supervenience-Representationism (SR) is the claim that every change in phenomenal character necessitates a change in the content of the experience, but not vice versa. But SR leaves open *which* change in content determines a specific change in character. In Ground-Representationism (GR), however, not any change will do: the change in character must have a change in content that *accounts* for the change in character.

Say you experience a change in the size of a gap, e.g. it grows larger. If that experience's character merely *supervenes* on its content, then the appearance of a growing gap does not necessitate that your experience is *about* a growing gap—*something* has to change in content, but it doesn't need to be this specific change. This appearance may be brought about by a change from being about a gap of size *x* to a smaller gap of size *y*, or about the gap changing color,

or about your toe starting to twitch while you look at the gap—any change might do without violating the letter of SR. However, if an experience's character is *grounded* in its content, then the change in content must account for the appearance. It seems that only being about a growing gap truly accounts for the appearance of a growing gap. If we want to be Representationists, GR seems like our best option: it allows us to (i) differentiate content from character, (ii) see content as more basic than character, (iii) capture that phenomenal character is dependent on content, but not *vice versa*, and (iv) make content accountable for character.

However, character is *not* grounded in content, Block argues:⁴ GR is false. This assessment is motivated by empirical considerations. There are many gems in Block's article, but I will focus mainly on the crown jewel, which is the argument based on "phenomenal precision". It is subtle and intricate, so my first step is to reconstruct it (with a bit of elaboration) in section 3. In section 4, I point to a few oddities and tensions I see with Block's other work. I do not see these tensions as offering a decisive blow to his argument, but as a plea for an elaboration on how Block thinks about phenomenal precision. (My main argument meanders through the main text. I keep it concise, but some points deserve some technical elaboration—thus the abundance of footnotes. They may be treated like beetroot on a buffet, i.e. skipped with clean conscience.)

2 "Phenomenal precision"

The notion of *phenomenal precision* plays an important role in Block's argument. He (this collection, p. 45 & 47) admits that it is a notion in need of clarification—but one where a lack of definition ought not give us headaches, since many concepts pertaining to phenomenality lack definability.⁵

aesthetic pictorial experiences. We may apply this distinction to experiences more generally.

² See also the article in this collection.

³ At least, most often explanations are seen as asymmetric. For example, Schindler (2013) has remarked that the mechanistic explanations à la Craver (2007) violate asymmetry, which he sees as a shortcoming of Craver's account.

⁴ This is not his first argument against Representationism (see e.g., Block 1996), but I will focus mainly on his *The Puzzle of Perceptual Precision* in this collection.

⁵ This has become somewhat like a signature move for Block. Consider e.g.: "You ask: What is it that philosophers have called qualitative states? I answer, only half in jest: As Louis Armstrong said when asked what jazz is, 'If you got to ask, you aint

We can think of precision as connected to bandwidth. What does that mean? Some variations in the external world do not factor into how the world feels to us. For example, one cannot differentiate a grating of 20% contrast from one of 20.2%, or a pain caused by heat of 480 millicalories per second per square centimeter from one caused by 640 mc/sec/cm² (Hardy et al. 1940; Hardy et al. 1952). However, there is a point where the variance in the stimulus becomes just noticeable, e.g., a pain of 660 mc/sec/cm² does feel different from one at 480. This can be measured behaviorally, namely if a subject is able to distinguish one item of type A from another of type B above chance based on the relevant feature (e.g., if 75% of all presented items are distinguished correctly). So all the variance that I cannot distinguish perceptually between two just noticeable differences (JNDs) is covered by percepts with the same phenomenal character. That is, if I have a percept *a*, different states of affairs may have caused *a*—and the phenomenal character of the percept does not convey its real cause. So percepts ought to count as a bit imprecise. The more cases are covered by a percept, the less precise: a visual-contrast-percept that can be caused by $20 \pm 1\%$ contrast is more precise than one that can be caused by $20 \pm 3\%$.

never gonna get to know.” (Block 1991, p. 217)

Or: “I cannot define [phenomenal consciousness] in any remotely noncircular way. I don’t consider this an embarrassment. The history of reductive definitions in philosophy should lead one not to expect a reductive definition of anything. The best one can do for [phenomenal consciousness] is in some respects worse than for many other concepts, though, because really all one can do is point to the phenomenon (cf. Goldman 1993a). Nonetheless, it is important to point properly.” (Block 1997, p. 230)

He continues by stating that synonyms and examples are the best way of conveying what is meant by “phenomenal consciousness”. I do not think that this is unreasonable, but instead intellectually honest. Chalmers (2011, p. 545) may have provided a good explanation of why it is so hard to provide a real definition for “phenomenality”: It might be a bedrock concept, which cannot be decomposed into more basic concepts, because it is itself most basic—it captures the fundamental distinction between reality and its appearance to us:

“[...] a dispute is bedrock relative to an expression: so the dispute over ‘Mice are conscious’ might be bedrock with respect to ‘conscious’ but not with respect to ‘mice’. A substantive dispute is bedrock relative to an expression *E* when no underlying dispute can be found by applying the method of elimination to *E* [i.e., replacing *E* in disputes by another expression where people agree on the meaning]: roughly, when there is no underlying dispute that does not involve *E* or cognates.”

If an expression is bedrock, then it cannot be elucidated by conceptual analysis—and there cannot be any non-stipulative real definition.

Percepts have representational and phenomenal aspects—content and character. Precision certainly makes sense when it comes to content, because “[t]he representational content of a perception is—constitutively—the veridicality conditions”, Block writes (this collection, p. 27).⁶ So we can look at the range of cases in the world that make a percept veridical, and thereby determine its degree of representational precision based on the range of cases that may have caused it in that obtaining condition. If, for example, a Gabor patch with 22% contrast looks just like one with a 28% contrast, then the representational content of this percept has a degree of precision of at least 6%, because all cases between 22% and 28% are covered by the same phenomenal appearance. Otherwise, these two Gabors would not look the same.

Representational precision makes sense—but how about *phenomenal* precision? Intuitively, phenomenal precision sounds good: things may appear *red* or *crimson*, and because all things crimson are a subset of all things red, the bandwidth of both ways of seeing-as differs—and therefore they ought to count as differently precise.

But if we can diagnose differences in the degree of phenomenal precision, we need a way of estimating its degree. How would we do this? GR provides an easy answer: phenomenal precision is grounded in representational precision, so we can use the same methods by which we estimate representational precision to estimate phenomenal precision. But in an argument where GR is under scrutiny, one cannot presume this without begging the question. So we must look for another way of estimating phenomenal precision.

For this purpose, Block suggests the Phenomenal Precision Principle (PPP), which we may reconstruct as: If the percept of item i_1 and the percept of item i_2 are phenomenally indistinguishable with respect to some feature *F* under condition *A*, but phenomenally determinately different vis-à-vis *F* under condition *B*, then the experience in *A* is less precise than in *B*.

⁶ This is in the spirit of Burge (2010, pp. 55–60), whom Block cites in this context.

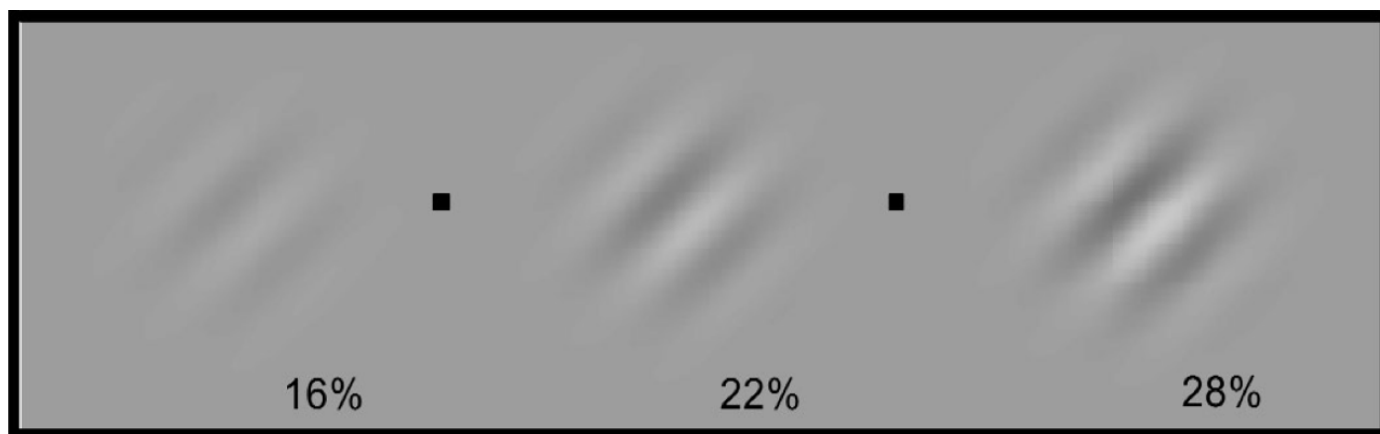


Figure 1: If one fixates on one of the black dots but actively attends to the lower-contrast patch to the left, the two patches to the right and left of the dot will appear alike. If one gazes freely or attends to the right, the difference in contrast is obvious or even more pronounced. Taken from Carrasco et al. (2004, p. 310).

So if I cannot differentiate two stimuli by their contrast in condition *A*, but can differentiate the two by contrast in condition *B*, then my experience in *A* is less precise than in *B*. Why? Because if I can tell the two items apart phenomenally, then I can distinguish cases, and therefore the bandwidth of that experience is narrower.

Block uses differences in phenomenal precision prominently in an argument against GR: he believes that in some cases, phenomenal precision (*p*-precision) and representational precision (*r*-precision) can fall apart. If GR were true, such that representational features must account for phenomenal ones, then this cannot be the case. But this is exactly what happens, according to Block: “there is evidence that attended and foveal perception can be greater in [phenomenal precision] without involving awareness of more precise environmental properties” (this collection, p. 41). Then, GR is false.

3 Block’s precision argument

What evidence speaks for Block’s thesis that “attended and foveal perception can be greater in [phenomenal precision] without involving awareness of more precise environmental properties” (this collection, p. 41)? (For those who have read the original article and have a firm grasp of the argument based on precision, this part may be skipped for the discussion in section 4.)

3.1 The stimulus and the conditions of viewing

Consider the stimuli in figure 1 taken from Carrasco et al. (2004, p. 310), and mentioned by Block twice (figure 7 and 9 in his article). It shows three Gabor patches of 16%, 22%, and 28% contrast—call these stimuli g_{16} , g_{22} , and g_{28} respectively. If we look directly (i.e., foveate) *at* and *attend to* each of these stimuli, the percepts they cause are decidedly different to each other. Call this condition “SFAG” for *stimuli foveated, attention on gabors*. However, if we fixate on the black spot between the patches (such that the patches are more in the periphery of our visual field) but *attend* to the one with lower contrast (i.e., to the left of where we fixate), then the percepts they cause appear indistinguishable from one another. Call this condition “SPAL” for *stimuli peripheral, attention on lower contrast*. This comparative indistinguishability does not arise if we attend to the higher contrast patch or to the spot in the middle. Call these conditions “SPA H ” for *stimuli peripheral, attention to higher contrast* and “SPA F ” for *stimuli peripheral, attention to fixation spot*, respectively.⁷

⁷ See also table 1. We may also introduce the following formalism: $Ch(x)$ stands for the character of a percept of x (the external stimulus); $Ch(x\&y)$ stands for the character of perceiving stimuli x and y together, i.e., a mereological fusion of simultaneously occurring characters at a moment in time t . The comparative character is then:

Table 1: The character in each condition of viewing/attending to the stimulus of 22% and 28% in figure 1. (See also footnote 7.)

Abbreviation	Condition	Character
SFAS	Stimulus foveated, attention on stimulus	distinguishable
SPAS	Stimulus peripheral, attention on fixation spot	distinguishable
SPAH	Stimulus peripheral, attention on higher contrast	distinguishable
SPAL	Stimulus peripheral, attention on lower contrast	indistinguishable

3.2 The evidence for attention influencing appearance

It seems that attention alters appearance. Our main evidence is introspective: we can reliably produce such changes in appearance from SPAH to SPAF to SPAL by shifting attention. This works even if we know of the effect.

To an external observer, there is evidence available from naïve subjects: If these subjects have to name the orientation of the patch with the higher contrast (\swarrow vs. \searrow), they choose at chance level in SPAL *even if there is a contrast difference of 6%*.⁸ (The shift in attention was

SFAG: $Ch(g_{22}) \neq Ch(g_{28})$

SPAF: $Ch(g_{22} \& g_{28}) \approx (Ch(g_{22}) \neq Ch(g_{28}))$

SPAH: $Ch(g_{22} \& g_{28}) \approx (Ch(g_{22}) \neq Ch(g_{28}))$

SPAL: $Ch(g_{22} \& g_{28}) \approx (Ch(g_{22}) = Ch(g_{28}))$

Note that just because phenomenal parts share identical phenomenal character, these parts themselves need not be identical: they may occur at different moments in time, be part of different phenomenal wholes, or be arranged in a different manner; if any of these extrinsic properties were among the identity conditions of phenomenal parts, then $a \neq b$; however, the character of some a may still be identical to the character of b .

If phenomenally-unified percepts are mereologically organized, as suggested by Bayne (2010), as well as Wiese & Metzinger (2012), then: if a percept at $t1$ has $Ch(x \& y)^{t1} = (Ch(x)^{t1} \neq Ch(y)^{t1})$ and a percept at $t2$ has $Ch(x \& y)^{t2} = (Ch(x)^{t2} = Ch(y)^{t2})$, then $Ch(x)^{t1} \neq Ch(x)^{t2}$ or $Ch(y)^{t1} \neq Ch(y)^{t2}$. This is the case in the experiment by Carrasco et al. (2004, p. 310); so under these presumptions, the appearances of parts of the overall percept must change between the conditions, because the character of the whole changes.

⁸ That is, the point of subjective equality (PSE) differs between the conditions. PSEs are determined by that configuration where a forced choice between stimuli is chancy. Carrasco et al. (see 2004, p. 311, figure 5a) kept a g_{22} fixed (standard) and varied the other patch (test). So one condition (cue to test) covered what I call here SPAH and SPAL, as in some cases, the test patch had a lower, in some a higher contrast than the standard g_{22} . (This is merely a difference in presentation, which does not influence the overall argument. Their presentation simply provides a continuous psychometric

exogenously triggered by a visual cue 27ms prior to stimulus onset.) Because subjects have to decide which grating *looks* higher in contrast, and pick the lower or the higher contrast patch at random, it is reasonable to assume that the two look the same: they have identical character in SPAL. Thus, attention affects appearance.

3.3 The contents and degree of r -precision in different conditions

So the character of comparative percepts (the character of experiencing two patches together) differs between these conditions, even if the stimuli and the way we fixate remain the same. But what about the respective *contents*?

In SFAG, we can clearly tell the patches apart. If percepts are constitutively veridical (because otherwise they are not percepts, but illusions or hallucinations), then the content of a percept is determined by the actual world. Thus, the content of each percept of a patch is (approximately) its actual contrast.⁹

In SPAF, the patches look different. However, as our ability to tell contrasts apart is a bit lower in the periphery, the contrast-JND is a bit higher—say, 3%.¹⁰ So the content of the comparative percept is one where the content of each percept is less precise, but still discernible from another: its actual contrast within the range of a peripheral contrast-JND.

In SPAH, the comparative contrast between the patches is more pronounced. We cannot explain this if the content in SPAH is the same as in

curve.) If the fixation spot was cued, the PSE reflects reality: a g_{22} looks most like g_{22} ; if the standard (g_{22}) was cued, the test patch had to have a higher contrast to look similar: a g_{28} looked most like a g_{22} ; if the test patch was cued, the uncued patch had to have a lower contrast to look most similar to the test: a g_{16} looked most like a g_{22} .

⁹ The actual content is a bit more imprecise, i.e., within the range of 1 foveal contrast-JND, which is roughly 1%. Block suggests 2% overall. In personal communication, Frank Jäkel estimated that (under ideal experimental conditions with optimal stimuli) the contrast-JND could be a log-unit lower than that: 1% provides a good ballpark estimate for many conditions. He based this estimation on his own work done for the study published in Jäkel & Wichmann (2006). See also Carney et al. (2000), Pelli & Bex (2013), and the *locus classicus*: Fechner (1860, pp. 150ff.).

¹⁰ See Banks et al. (1991, p. 1779). Although they do not specifically mention JNDs, they do provide data about contrast sensitivity in different degrees of peripheral eccentricity, which suggests some increase: “[T]he ideal [contrast sensitivity functions] do not exhibit the large contrast sensitivity losses that one observes in humans with increasing eccentricity.”

SPAF. Somehow, the contents ought to differ more than in SPAF. One way to do this is to see one as more r -precise than the other. Then it is easier to tell the two apart, because there is no content-overlap. Another way would be to assume that one becomes less r -precise. Then it is easier to tell them apart because the respective minima and maxima are further apart.¹¹

Table 2: The content and its degree of precision in each condition of viewing/attending to the stimulus of 22% and 28% in figure 1. (See also footnote 12.)

Condition	Content + Bandwidth estimation
SFAS	% of actual contrast ± 1 foveal JND _{attended} ($\sim 1\%$);
SPAS	% of actual contrast ± 1 peripheral JND _{unattended} ($\sim 3\%$);
SPAH	% of actual contrast ± 1 peripheral JND _{attended} / ± 1 peripheral JND _{unattended} ; (my estimation: $\sim 2\% / \sim 3\%$; see also footnote 11)
SPAL	at least the open interval between actual contrasts, here $\geq 6\%$;

In SPAL, the comparative percept (g_{22} & g_{28} together) is such that the two patches are indistinguishable. So our percept is strictly speaking non-veridical. In order to make it veridical, one has to assign a quite imprecise content: it must at least cover both actual contrasts—i.e., be greater or equal to the interval that includes the actual contrasts as endpoints: $[22\%, 28\%]$.¹²

¹¹ Block might argue that we lack a principled reason to choose one over the other as being more or less imprecise. The argument mirrors the one he gives concerning veridicality (see Block this collection, pp. 26ff.). As veridicality determines the contents of percepts, one can easily adapt it: intuitively, one might think that the patch one attends to is more veridical; but attention changes appearance, so the unattended one might be more veridical; but as one mostly acts on what one attends to, it would be advantageous if what one acts on was most veridical. So we are stuck in a rut. The comparative percept in SPAL is illusory, but *as a percept*, it must be (partially) veridical. Block's suggestion is (or ought to be) that we should assume that each is veridical, but less r -precise. I'd agree. But I think we can do more: when we focus on the higher contrast Gabor, this increases the distance in r -precision between the compared percepts, and thereby ought to render them more discernible. If so, then this might apply to SPAL as well, such that the one we attend to is more precise. I pick up on this in footnote 12.

¹² See table 2. More formally, let $Co(x)$ stand for the content of our percept of x , and $Co(x \& y)$ for the content of the comparative percept of x and y together. Given the external content-determination of percepts and our understanding of JNDs, we can be a bit more precise about how imprecise content is in the different conditions.

SFAG: $Co(g_{22}) = 22 \pm \sim 1\%$; $Co(g_{28}) = 28 \pm \sim 1\%$

SPAF: $Co(g_{22} \& g_{28}) \approx (Co(g_{22}) \geq 22 \pm \sim 3\%, Co(g_{28}) \geq 28 \pm \sim 3\%)$

SPAH: $Co(g_{22} \& g_{28}) \approx (Co(g_{22}) \geq 22 \pm \sim 3\%, Co(g_{28}) \geq 28 \pm \sim 2\%)$

SPAL: $Co(g_{22} \& g_{28}) \approx (Co(g_{22}) \geq [22\%, 28\%], Co(g_{28}) \gg [22\%, 28\%])$

3.4 Estimating the degree of p -precision in the different conditions

So we know the percepts' contents and r -precision in the different conditions—but how about their p -precision? Block agrees that this is hard to estimate correctly. But the PPP gives us a rough guide: if the percept of item i_1 and the percept of item i_2 are phenomenally indistinguishable with respect to some feature F under condition A , but phenomenally determinately different vis-à-vis F under condition B , then the experience in A is less precise than in B vis-à-vis F . However, the case becomes more complicated, because we also have to think of the p -precision of *comparative percepts* (experiences as a whole) in addition to the *percepts compared* (the parts of whole experiences), akin to what we did in the case of r -precision.

3.4.1 Perceptual wholes and perceptual parts

At each moment, you have a broad range of different sensations; but all of these together are parts of one massive phenomenal *me-here-now-with-this-and-that-whole*: at a bar, you smell the mixture of spilt beer and sweat, taste the medicinal-peaty taste of your Lagavulin, while you ogle a lovely co-member of your species—who makes you feel your heart pumping in your chest. But you don't feel all these separately; they are fused into one fleeting holistic experience.

If phenomenal wholes are not character-identical, there must be a difference in their parts; but some distinguishable phenomenal wholes may still share parts with identical phenomenal character: the feel of your beating heart while ogling may be phenomenally identical to the feeling of your beating heart after escaping the oglee's significant other.

3.4.2 Unattended parts can share character with attended parts

Just as temperature can alter the taste of sugar to caramel without being sugar or caramel, attention can affect phenomenal character without itself having a phenomenal character: attention *alters* the appearance of x , but there seems to be no additional phenomenal character

as of *attending to x*. If so, then the phenomenal character of a perceptual part itself does not determine whether this part is attended to or not.¹³ So a percept that is now in the attentional limelight may share its character with a counterpart in the attentional shadow: If I attend to a leaf in a tree, the leaf I focus on may look just as green as a leaf in my visual periphery that I experience but don't care about. This is one interpretation of SPAL: the percept of the attended peripheral g_{22} shares its character with the percept of the unattended peripheral g_{28} -patch.¹⁴

3.4.3 An estimation of p -precision in the different conditions

Now, we may consider what the p -precision is in our cases. In SFAG, over the range of 1 foveal contrast-JND, all percepts look the same. This is the most p -precise that the character of a percept can be. The p -precision range is then roughly centered around some value $n \pm x\%$, where x is approximately 1 foveally attended contrast-JND.¹⁵

In SPAF, the patches look determinately different; and in SPAH, they look even more dif-

ferent. It is in the spirit of PPP (see p. 3) that the comparative percept in SPAH is more p -precise than the comparative percept in SPAF.

But because the character of a percept is independent of whether one attends to or foveates on it, each compared percept (the parts of which the comparative percept is composed) ought to be similarly p -precise as in SFAG: if parts inside and outside the focus of attention can share phenomenal character, and if this holds for all characters, then the same range of characters can appear anywhere in our visual experience. So we ought to expect the same range of PPP-cases in the periphery as in the fovea. Then, the character ought to count as similarly p -precise.

3.5 The argument

If I am correct so far, we can state the following: (P1) If the character of an attended and an unattended percept can be identical (section 3.4.2), then perceptual parts are overall more p -precise than r -precise, because the range of p -precision-values of compared percepts is stable in all conditions (table 3), but the range of r -precision must vary in order to account for the veridicality of percepts (table 2). (P2) If the character of an attended and an unattended percept can be identical, then our compared percepts (the parts of the comparative percept) are more p -precise than r -precise in SPAL.¹⁶ But if GR were true, then there must be a representational feature that accounts for each phenomenal feature. This applies to precision as well, because—according to Block— p -precision is a phenomenal feature of one's perception. So if GR were true, representational precision must account for phenomenal precision. But (P1) and (P2) stand in direct opposition to this. So, by *modus tollens*, GR ought to be considered false.

4 On the notion of “phenomenal precision”

Any argument against Representationism has an initial appeal to me. Ned Block's is at the cutting edge of empirical research and subtle in its argu-

¹³ One reviewer doubted whether this holds generally. It might be that a weaker version is easier to defend: an appearance does not *necessarily* specify whether it is attended to or not. I suspect that Block tends towards a stronger reading, as it seems to be in line with the dissociation between phenomenal consciousness and access. My reconstruction hinges on the strong version. For otherwise, the stability of p -precision I suggest in section 3.4.3 does not arise. So if there are good reasons to doubt the strong version, there are good reasons to doubt my reconstruction ↓ and also Block's argument itself, I believe. Here, I simply admit this weakness, but cannot follow up on this criticism due to lack of space. However, I see no good reason for assuming that attention has a unique and distinguishable perceptual character. (It might even be contentious whether it has a cognitive, active, or any phenomenal character at all, but I will not get into this here.)

¹⁴ It is unclear whether the identity in character must be part of the experience for Block's argument. It seems that he thinks this way. But consider Williamson (1990, p. 60; my emphasis), who writes that the “discriminability of a pair of characters as presented by a pair of experiences depends on *non-qualitative* relations between the experiences — relations not fixed by the way in which the experiences present their characters — which facilitate or hinder discrimination [...]” He envisions *where* the compared characters are placed in time and the visual field, but one might also consider, as I do in section 4.3, that our ability or inability to tell characters apart is a dependent on our cognitive abilities.

¹⁵ We cannot give the exact value of n , because the character of a percept is independent of whether one attends to it or foveates on it; and in SPAL, percepts of different actual contrast can share the same character. So we cannot associate the p -precision value with any value pertaining to a stimulus. Still, we may assume that it has a value. So I use some mock-value n .

¹⁶ Block's main argument rests on (P2)↓ but I hope that (P1) is in his spirit.

mentation. But I suspect that its crown jewel, “phenomenal precision”, has a few shady facets.

“Phenomenal precision”, Block admits, is in need of clarification. The guiding example for Block ([this collection](#)) is where “[t]he experience of a color as red is less phenomenologically [sic] precise than the experience of a color as crimson”. Here I want to focus a bit more on how we may understand p -precision, what it might and what it ought not to mean in the context of Block’s work.

Table 3: The approximated phenomenal precision in each condition of viewing/attending. How p -precise the comparative percepts ($g_{22}\&g_{28}$ *perceived together*) are can be ordered from lowest to highest: SPAH>SPAF>SPAL. (See also footnotes 7 and 12.)

Condition	Approximated Phenomenal Bandwidth
SFAG	$n_i \pm x\%$ of contrast, where x is roughly 1 foveal JND; $\text{Ch}(g_{22\%}) = n_1 \pm x\%$; $\text{Ch}(g_{28\%}) = n_2 \pm x\%$
SPAF	$\text{Ch}(g_{22\%}\&g_{28\%}) =$ $((\text{Ch}(g_{22\%}) = n_3 \pm x\%) < (\text{Ch}(g_{28\%}) = n_4 \pm x\%))$
SPAH	$\text{Ch}(g_{22\%}\&g_{28\%}) =$ $((\text{Ch}(g_{22\%}) = n_5 \pm x\%) \ll (\text{Ch}(g_{28\%}) = n_6 \pm x\%))$
SPAL	$\text{Ch}(g_{22\%}\&g_{28\%}) =$ $((\text{Ch}(g_{22\%}) = n_7 \pm x\%) = (\text{Ch}(g_{28\%}) = n_4 \pm x\%))$

4.1 Lower bounds of p -precision

The way I estimated phenomenal precision in my reconstruction was as follows: consider, first, how a controlled stimulus appears under ideal conditions (e.g., rested, attending, etc.) as some phenomenal feature. For example, how *blue*₃₄ in a standardized patch looks *as blue*, how an olfactory sample (e.g., a *CAS 93686-30-7, Ext. Sup. I, 1000ppm*) smells *as Ylang-Ylang*, how 480 mc/sec/cm² feels *as pain*, and so on. Then, see how much variance in the stimulus is *not* mirrored in the appearance as F : for example, the pain caused by 480 mc/sec/cm² is not reliably discernible from one caused by 640 mc/sec/cm²; instead both feel *as pain near maximal intensity*. Because I cannot differentiate between 480 mc/sec/cm² and 640 mc/sec/cm² by the feeling they cause, my pain feeling’s p -precision must at least cover these values. This provides us with a lower bound

for that specific feeling of pain. More broadly, for any phenomenal character—i.e., experiencing something as F (e.g., a color as red, a tone as $C\sharp$, a patch as having 28% contrast)—, the lower bound of its p -precision is that range of cases one cannot distinguish by experiencing as F under ideal conditions. That’s what the PPP suggests (see p. 3).

This allows us to make sense of the *red vs. crimson* example: *crimson* is a very specific phenomenal feel, which allows for very little variation while remaining crimson. *Red*, on the other hand, allows variation along the whole spectrum, from coral and vermillion, via crimson and oxblood, to maroon. So any experience of a color as *crimson* is also likely to be¹⁷ an experience of a color as *red*—but so is an experience of a color as *vermillion*. And the range of cases that may cause an experience of red compared to those that may cause an experience of crimson under ideal condition is larger. So experiencing as crimson is more p -precise than experiencing as red.

4.2 A need for solely generic phenomenology?

I think that triviality or contradiction looms if we do not add another constraint to be satisfied: In order for p -precision to be non-trivial, there must be the possibility of experiencing a color as red, but *not* as crimson₄₂, vermillion₁₁, coral₁₉, oxblood₈₁, etc. That is, there must be a way of experiencing something as a higher-order property \mathfrak{F} , *without* experiencing it as any first-order property F_1, F_2, \dots subsumable under \mathfrak{F} . Rick Grush (2007) has called this *Generic Phenomenology*—but I am speaking more specifically of *solely generic phenomenology* (SGP), i.e., generic phenomenology without an accompanying and subsumable concretum.¹⁸

Why ought we commit ourselves to SGP? Because otherwise the p -precision of an experience

¹⁷ Likely but not necessarily, because *experiencing as crimson* is not necessarily related to *experiencing as red*. Conceptual or nomological relations do not necessarily transfer to the realm of experiences. Imagine seeing an animal *as a mouse*. One does not thereby see it *as an owner of a heart*, or *as a member of the phylum chordata* even though all mice belong to each category necessarily.

¹⁸ See also the discussion and specifically Block’s response R2 on Block (2007) for more on generic phenomenology.

is either contradictory, generally minimal, or generally maximal, which trivializes the notion. Why?

Look at the color in figure 2. What is the p -precision of this color impression?



Figure 2: A stimulus of color 660000 or 16-86-94-42 CMYK.

You probably experience this color as red, but also as having a specific shade of red—for which you might lack a name, but let us baptize it *cayenne*₆₆. It is natural to assume that this holds for all color impressions, e.g., that whenever you experience a color as red, you also experience it as a most specific shade. In this case, there is no SGP—just a generic phenomenology accompanied by concrete and “subsumable” phenomenology. What might be the p -precision of your color impression in this scenario?

You might think that this color experience has two p -precision values:¹⁹ The first value is for being experienced as red, and the second for being experienced as *cayenne*₆₆. But this seems contradictory: why should one and the same experience of a color have two p -precision values, but only one for r -precision? And for that matter, why not three values for p -precision? You likely experience the color not only as *cayenne*₆₆ and as red, but also *as a color*? Why not four, then, if you experience it *as a visual experience*? Or five, if you experience it *as something*? Or even six, if you experience it *as phenomenal*? The more options we consider, the less sense it makes to speak of *the* p -precision of a percept at all. But this is needed for Block’s argument, where changes in p -precision are lower than the respective changes in r -precision. This hardly works if we allow multiple values. So we should assign experiences only one p -precision value.

If we have to assign this color experience only *one* p -precision value, we could either choose

the lowest or the highest feature. Either option looks arbitrary, which is already bad. But it gets worse if we reject SGP: If we chose the lowest feature (*cayenne*₆₆) and if there must always be a lowest feature, then all experiences of a type have the same level of p -precision *and* this value must be stable. If we chose the higher feature (red), then there is no reason to stop there: we certainly experience the color as a color, as a visual impression, or as *something*. But the character of *being something* applies to (almost) every experience. So all experiences would again be equally and fixedly p -precise. Both cases seem to trivialize the notion of p -precision, because it always stays the same. So p -precision seems either contradictory or static and trivial without SGP.

We ought to accept SGP in order to allow for variance in p -precision: we can experience some color as red, but *not* as e.g., *cayenne*₆₆ and so on. More generally, we can experience something *only* as a higher-order property \mathfrak{F} *without* experiencing it as any lower order property F_1, F_2, \dots subsumable under \mathfrak{F} . Then, different experiences allow for different degrees of indeterminacy and therefore different degrees of p -precision.

However, SGP has been introduced to argue *against* Block: According to Grush (2007), if we accept the possibility of generic phenomenology, then we could see something as *some* letter without seeing it as *a specific* letter (A, B, ...). This affects one’s interpretation of the Sperling experiment: Sperling (1960) showed participants a grid of letters, which they identified *as letters* from the short impression they got. Yet they could not identify and recount all of them. But when they were cued to repeat a specific line by a tone *after* the stimulus disappeared, they were able to recount the letters in that line without fault. Block (2007) has used this and other experiments to argue that phenomenality goes beyond what we can cognitively access: people have a full phenomenal impression, but cannot access all the information available in their experience. Their experience is concrete, but their introspective access is shaky. SGP proponents counter that one can have generic experiences while all the underlying concrete information is subconscious. So before the cue, subjects experience concretely according to Block, but generically according to

¹⁹ I speak as if we could know the determinate value of p -precision given as a real number here. But this is not required: There could be a determinate value without us being able to know it.

proponents of SGP. Allowing SGP thus blocks Block.

Additionally, generic phenomenology seems to be closely associated with *symbolic* or *rule-based*²⁰ representation. *Imagistic* representation, on the other hand, does not allow for such indeterminacy, because images exploit the isomorphisms between concreta. I can write “The cat is on the mat” without saying anything about whether the cat is a Siamese or a Maine Coon, or whether the mat is filled with feathers or made of bamboo, or whether the cat reclines, sits, or scratches on the mat. The sentence can represent the fact without resembling a cat or a mat at all. However, if I want to represent the fact that *the cat is on the mat* in an image, I have to depict something concrete: a specific cat at some position on a mat doing something. The common understanding of images is that they are concrete and as such determined in all their lowest-order properties. Analogue representations more generally exploit concreteness in order to represent by isomorphism.²¹

Introspectively, our phenomenal experiences resemble images. If phenomenal experience represents imagistically, then there cannot be SGP—and p -precision seems dangerously close to being trivial; if phenomenal experience is non-imagistic, then we can allow for SGP and render p -precision non-trivial—but this is in tension with some of Block’s other work and our introspective evidence.

Maybe a fixed p -precision value need not be bad for Block’s argument: if the p -precision of percepts is fixed, but r -precision varies, then there is a phenomenal feature that is not grounded in a representational feature. Thus, GR is false. However, Representationist have an easy reply: GR does not claim that all changes in representational features must be mirrored in phenomenal features; representational features only need to account for phenomenal features. If p -precision is fixed, then it might be grounded in

there being r -precision *at all*. For Block, accepting SGP might be a good option here—but not elsewhere.

4.3 Introspective imprecision?

There might be a way to reject SGP, but still account for our belief that we can experience a color as red without experiencing it as crimson. Maybe *experiencing as \mathfrak{F} without experiencing as any subsumable F* does not apply to phenomenal experiences, but to our *access* to them. That is, maybe there is *introspective* rather than phenomenal precision.²² This might go along the lines of Block’s interpretation of the Sperling experiment: we experience very specific shapes, but introspectively, we are only able to label them as *letters*, not as *A*, *B*, etc. So maybe the phenomenal aspects of our experience have fixed precision because it is never solely generic; but our introspective judgements are not fixed in precision because we can introspect some experience solely generically. That is, we may *judge* an experience of cayenne₆₆ to be red although we actually experience it as cayenne₆₆. If perception can be more or less imprecise, why can’t “internal perception”?²³

The notion of introspective imprecision, however, is not easily applied to the example of the patches in Carrasco et al. (2004): if we introspect on their appearance, then we judge them not as imprecise, but as of precisely the same contrast.

But maybe the imprecision of introspective access is not itself introspectively available: Our introspective access might be limited, such that all we can tell is that the patch

²² A reviewer noted that the limits of our vocabulary and our verbalization skills more generally might account for the lack of discrimination skills just as well as introspective imprecision. Even though this is a valuable point, I do not develop it here. First because I want to stick as close to the occurrent percepts as possible, not to our cognitive grasp of such percepts; second because introspective access precedes verbalization of the introspected; and third because failures of verbalization do not account for “introspective data”. Block is unapologetic about taking introspection seriously; a supportive critic should take it seriously as well.

²³ Many philosophers liken introspection to a form of perception, e.g., Locke (2008, II.xxvii.§9 & II.i.§4), Kant (2008, AA,III, tr.Äst.,§2), Brentano (1874), James (1890), Boring (1953, p. 170), Armstrong (1980, p. 61), Lycan (1996, p. 334), and to some degree Goldman (2006, pp. 242ff.) as well as Churchland (1985, 2005).

²⁰ The rule might be cultural, as with language, but also natural, as with causes: that the word “red” means the color ■ is based on a cultural rule; that *smoke* means *fire* is based on a natural rule.

²¹ This case has been made by Kosslyn (1980, p. 31) as well as Gombrich (2002). But see also Haugeland (1981), Lewis (1971), and Jackson (1960).



Figure 3: A 24-bit picture compared to a 6-bit picture. Shifting attention between the two while focusing on the + ought to provide some experience of indistinguishable character if experience in the periphery is more coarse grained.

we attend to is like the patch we don't attend to *in some respect*. However, our general introspective bias—that we think ourselves as authoritative about our own minds leads us to overrate what introspection offers: what we introspect as being more or less alike is judged as being strictly alike. This bias towards seeing ourselves as introspectively authoritative independent of whether we introspected successfully or not might lead to a wide variety of false beliefs about phenomenality.

The upshot would be that introspective imprecision is compatible with Block's distinction between access- and phenomenal consciousness. But introspective imprecision leaves it open whether SGP holds or not. It seems that we cannot decide based on introspection whether the character of our percepts or our introspective access to them is imprecise. We would need some other access to our phenomenality in order to settle the issue; but at this moment in time, nothing comes to mind that offers decisive evidence.²⁴

²⁴ This differs from the argument Block discusses: the attentional effect could be perceptual and conscious, but it is not really accessible what or how much actually changes in these circumstances due to introspective imprecision.

Block's writing suggests that he rejects introspective imprecision in this article (although he ought to accept it when defending the distinction between access- and phenomenal consciousness). If we reject it with him, how can we save the idea of percepts being more or less *p*-precise?

4.4 Limitation on characters?

The idea that parts of perceptual wholes can be more or less imprecise seems to stand in tension with the idea that all appearance-features can turn up anywhere in the phenomenal field: any appearance of contrast may appear in the fovea or periphery or where I attend or don't attend, etc. This had the odd consequence in my reconstruction that all phenomenal parts have the same degree of *p*-precision. How might we avoid this?

We could assume that the range of characters in the focus of attention and in the fovea is most fine-grained. Imagine being able to experience 100 shades of crimson in the attended fovea, but only 20 shades of crimson in the unattended periphery. This is reasonable for

contrasts as our sensitivity to it declines with eccentricity (Banks et al. 1991).

But this suggests that our experience is less continuous in the periphery. Instead, it is stepwise. This fits to the idea of precision having to do with bandwidths: in the attended fovea, we experience with a higher bit-rate than in the unattended periphery. It is like seeing a picture in 24- instead of 6-bit color depth (see figure 3).²⁵

But the coarse-grained character of experience outside of attention is not introspectable: if a light slowly changes color in our periphery, it does not look like it is doing so stepwise. It looks smooth and continuous. So somehow this idea only makes sense if we add the idea of introspective imprecision—and thereby inherit its problems.

So it is open how we should marry the idea of variances in phenomenal precision of a specific character with Block's overall view of conscious experience. Some more elucidation would be highly appreciated.

5 Conclusion

Ned Block has provided a beautiful argument against Ground Representationism—the position that for each phenomenal feature there is a representational feature that accounts for it. At its core is the notion of “phenomenal precision”: if we accept it, it seems that the degree of phenomenal precision of a percept changes differently to its degree of representational precision. Thus, there is no representational feature that accounts for this change in phenomenality—and Ground Representationism is false.

I have suggested a way of estimating phenomenal precision based on the assumption that parts of perceptual wholes can share characters independently of where they occur in the perceptual whole, and on the notion of a just noticeable difference as a lower bound of p -precision, which is inspired by Block's Phenomenal Precision Principle. Understood in this way, the argument shows what it is supposed to show: Ground Representationism is false.

But a deeper look at the notion of phenomenal precision suggests some tension with

Block's other work or with introspective evidence, which Block takes seriously. In order to allow for variation in the degree of precision, we have to accept that some of our experiences are not concrete, but solely generic. Such “solely generic phenomenology”, however, is a position mainly held by opponents to Block's *Access- vs. Phenomenal Consciousness-distinction*. Without accepting solely generic phenomenology, however, phenomenal precision seems either trivial (there is no variation) or contradictory (a percept can simultaneously have various degrees of p -precision). So the argument against Ground Representationism either hinges on a trivial or self-contradictory notion, *or* it is incompatible with Block's positions elsewhere. Patching this problem by allowing a limited range of characters outside our attention is again at odds with Block's other writing and with introspective evidence.

What is needed is a better understanding of phenomenal precision. What is it? How can we estimate it? I have suggested some possible ways to answer these questions, but all that I could come up with seems at odds with what Block has in mind. This certainly does not mean that there cannot be a suitable version of phenomenal precision that avoids these pitfalls—I am just unable to find it, and all that I can construct somehow goes against Block. I hope that Block has some ace up his sleeve, because the notion of phenomenal precision appears too fruitful to be abandoned too hastily.

Acknowledgements

I am very grateful to Frank Jäkel for help and discussion concerning the psychophysical side of this article, and to Thomas Metzinger as well as two anonymous reviewers for their supportive comments and criticism.

²⁵ The difference is one of 16.777.216 to 64 different colors.

References

- Armstrong, D. M. (1980). *The nature of mind and other essays*. Ithaca, NY: Cornell University Press.
- Banks, M. S., Sekuler, A. B. & Anderson, S. J. (1991). Peripheral spatial vision: Limits imposed by optics, photoreceptors, and receptor pooling. *Journal of the Optical Society of America A*, 8 (11), 1775-1787. [10.1364/JOSAA.8.001775](https://doi.org/10.1364/JOSAA.8.001775)
- Bayne, T. (2010). *The unity of consciousness*. Oxford, UK: Oxford University Press.
- Block, N. (1991). Troubles with functionalism. In David Rosenthal (Ed.) *The nature of mind* (pp. 211-228). Oxford, UK: Oxford University Press.
- (1996). Mental paint and mental latex. *Philosophical Issues*, 7, 19-49. [10.2307/1522889](https://doi.org/10.2307/1522889)
- (1997). On a confusion about the function of consciousness. In N. Block, O. Flanagan and G. Güzeldere (Eds.) *The nature of consciousness: Philosophical debates* (pp. 375-415). Cambridge, MA: MIT Press.
- (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences*, 30 (5-6), 481-548. [10.1017/S0140525X07002786](https://doi.org/10.1017/S0140525X07002786)
- (2015). The puzzle of perceptual precision. In T. Metzinger and J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- (forthcoming 2015). The Canberra Plan neglects ground. In T. Hogan, M. Sabates and D. Sosa (Eds.) *Qualia and mental causation in a physical world: Themes from the philosophy of Jaegwon Kim*. Cambridge, UK: Cambridge University Press. <http://www.nyu.edu/gsas/dept/philo/faculty/block/papers/Kimfestschrift.pdf>
- Boring, E. (1953). The history of introspection. *Psychological Bulletin*, 50 (3), 169-189. [10.1037/h0090793](https://doi.org/10.1037/h0090793)
- Brentano, F. (1874). *Psychologie vom empirischen Standpunkt*. Leipzig, GER: Duncker & Humblot.
- Burge, T. (2010). *Origins of objectivity*. Oxford, UK: Oxford University Press.
- Carney, T., Tyler, C. W., Watson, A. B., Makous, W., Beutter, B., Chen, C.-C., Norcia, A. M. & Klein, S. (2000). Modelfest first year results and plans for year two. In B. E. Rogowitz and T. N. Pappas (Eds.) *Human vision and electronic imaging V* (pp. 140-151). The International Society for Optical Engineering.
- Carrasco, M., Ling, S. & Read, S. (2004). Attention alters appearance. *Nature Neuroscience*, 7 (3), 308-313. [10.1038/nn1194](https://doi.org/10.1038/nn1194)
- Chalmers, D. J. (2011). Verbal disputes. *Philosophical Review*, 120 (4), 515-566. [10.1215/00318108-1334478](https://doi.org/10.1215/00318108-1334478)
- Churchland, P. M. (1985). Reduction, qualia, and the direct introspection of brain states. *The Journal of Philosophy*, 82 (1), 8-28.
- (2005). Chimerical colors: Some phenomenological predictions from cognitive neuroscience. *Philosophical Psychology*, 18 (5), 527-560.
- Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford, UK: Oxford University Press.
- Fechner, G. T. (1860). *Elemente der Psychophysik*. Leipzig, GER: Breitkopf & Härtel.
- Goldman, Alvin I. (2006). *Simulating minds*. Oxford, UK: Oxford University Press.
- Gombrich, E. (1982/2002). The visual image. *The image and the eye: Further studies in the psychology of pictorial representation* (pp. 137-161). London & New York: Phaidon.
- Grush, R. (2007). A plug for generic phenomenology. *Brain and Behavioral Sciences*, 30 (5/6), 504-505. [10.1017/S0140525X07002841](https://doi.org/10.1017/S0140525X07002841)
- Hardy, J. D., Wolff, H. G. & Goodell, H. (1940). Studies on pain. A new method for measuring pain threshold: Observations on spatial summation of pain. *The Journal of Clinical Investigation*, 19 (4), 649-657. [10.1172/JCI101168](https://doi.org/10.1172/JCI101168)
- Hardy, J. D., Wolff, H. G. & Goodell, H. (1952). *Pain sensations and reactions*. New York, NY: Hafner.
- Haugeland, J. (1981). Analog and analog. *Philosophical Topics*, 12 (1), 213-225.
- Jackson, A. S. (1960). *Analog computation*. New York, NY: McGraw-Hill.
- James, W. (1890/1957). *The Principles of Psychology*.
- Jäkel, F. & Wichmann, F. A. (2006). Spatial four-alternative forced-choice method is the preferred psychophysical method for naïve observers. *Journal of Vision*, 6 (11), 1307-1322. [10.1167/6.11.13](https://doi.org/10.1167/6.11.13)
- Kant, I. (1787/2008). *Kritik der reinen Vernunft*. Essen, GER: Korpora.
- Kosslyn, S. M. (1980). *Image and mind*. Cambridge, MA: Harvard University Press.
- Lewis, D. (1971). Analog and digital. *Nôus*, 5 (3), 321-327.
- Locke, John (1690/2008). *An essay concerning human understanding*. Oxford, UK: Oxford University Press.
- Lycan, W. G. (1996). *Consciousness and experience*. Cambridge, MA: MIT Press.
- Nanay, B. (2005). Is twofoldness necessary for representational seeing? *British Journal of Aesthetics*, 45 (3), 248-257. [10.1093/aesthj/ayi034](https://doi.org/10.1093/aesthj/ayi034)

- Pelli, D. G. & Bex, P. (2013). Measuring contrast sensitivity. *Vision Research*, 90 (0), 10 - 14. <http://dx.doi.org/10.1016/j.visres.2013.04.015>. <http://www.sciencedirect.com/science/article/pii/S0042698913001132>
- Schindler, S. (2013). Mechanistic explanation: Asymmetry lost. In D. Dieks & V. Karakostas (Eds.) *Recent Progress in Philosophy of Science: perspectives and foundational problems*. Dordrecht, NL: Springer. http://philsci-archive.pitt.edu/9557/1/Craver_symmetry_final.pdf
- Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied*, 74 (498), 1-29. [10.1037/h0093759](https://doi.org/10.1037/h0093759)
- Wiese, W. & Metzinger, T. (2012). Desiderata for a mereotopological theory of consciousness. In S. Edelman, T. Fekete and N. Zach (Eds.) *Being in time: Dynamical models of phenomenal experience* (pp. 185-209). Amsterdam, NL: John Benjamins.
- Williamson, T. (1990). *Identity and discrimination*. Oxford, UK: Blackwell.
- Wollheim, R. (1987). *Painting as an art*. Thames & Hudson Ltd.

Solely Generic Phenomenology

A Reply to Sascha Benjamin Fink

Ned Block

If representationism is true, phenomenal precision is given by representational precision. But what if representationism is false as I claim? Can we make sense of phenomenal precision? Fink argues that there is a danger of trivialization of phenomenal precision and that the one way out may be incompatible with my view that consciousness overflows cognition. I try to say more about how to clarify phenomenal precision and its relation to my views on overflow.

Keywords

Generic | Phenomenal precision | Phenomenology | Solely generic phenomenology | Specific

Author

[Ned Block](#)

ned.block@nyu.edu

New York University

New York, New York, U.S.A.

Commentator

[Sascha Benjamin Fink](#)

sfink@ovgu.de

Otto von Guericke Universität

Magdeburg, Germany

Universität Osnabrück

Osnabrück, Germany

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität

Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University

Melbourne, Australia

1 Introduction

I am grateful to Sascha Benjamin Fink for a thoughtful and insightful critique ([Fink 2015](#)) of my article ([Block 2015](#)). Fink's critique is full of novel and interesting ideas, formulations and proposals but is far too rich for me to respond to everything. I will focus on Fink's arguments to the effect that the concept of phenomenal precision is defective because there will be no unique precision to a phenomenal experience, specifically that phenomenal precision is either contradictory or trivialized by a "minimal" or "maximal" interpretation. I think Fink is right to focus on the concept of phenomenal precision since as he says

it is the aspect of my paper that most needs clarification. I argue that the key to solving the problem that Fink raises is to ask what the representationist should say about it. I then argue that the anti-representationist can make a similar move. In the last section I consider some variants of Fink's proposal for how to clarify phenomenal precision.

2 The thesis of solely generic phenomenology

I will start with the SGP thesis in Fink's terminology—the thesis that there can be "solely

generic phenomenology”. An example would be the experience of something as red without an experience as of any specific shade of red. Fink says I am forced to accept solely generic phenomenology but that it “has been introduced to argue *against* Block” (p. 10).

Fink is talking about my “overflow” arguments. These arguments are based partly on an experiment by George Sperling (Sperling 1960) that is covered in all respectable introductory psychology courses and on experiments from Victor Lamme’s lab in Amsterdam that combine Sperling with “change blindness” (Lamme 2003). The upshot of these lines of research is that there is more “capacity” in phenomenology than in cognition—or so I have argued. In the Sperling experiment, subjects are presented briefly with an array of letters, for example, 3 rows of 4 letters. Subjects say they see all or almost all of the letters but when asked to name the letters they saw after the stimulus has disappeared they can name only 3 or 4. Sperling’s innovation was to “cue” subjects to report one specific row after the offset of the stimulus, using a tone to indicate the row. The finding is that subjects can report 3 or 4 items from any given row, supporting the idea that their phenomenal “iconic” representation really did include information about the specific shapes of all or almost all the letters in the array.

In Lamme’s experiments the subject is briefly shown an array of, for example, 8 oriented rectangles. After the array is turned off, the screen goes dark for up to 4 seconds, then there is a second array of 8 rectangles in which one of the rectangles may have a different orientation (e.g., vertical rather than horizontal). A cue—a line pointing to the location of one of the rectangles—can occur when the screen is dark or, alternatively, when the second array appears. The subject’s task is to say whether the rectangle in the cued location has changed orientation. If there is no cue or if the cue comes after the new array has overwritten the iconic representation of the first array, subjects have a capacity of about 4 items. But subjects say they have a kind of image of the array in the dark period after the stimulus has gone off. When the cue is presented in the dark period

(up to 4 seconds later in some versions) after the stimulus has gone off they have a capacity of up to 7 of 8 items. I have argued that this pattern of results indicates that subjects have a persisting conscious mental iconic or imagistic representation of 7 of the 8 rectangles of sufficient specificity to compare orientations of the initial rectangles with the final display of rectangles even though they can “cognize” only about 4 of them—in the sense of storing them briefly in “working memory”. The upshot according to me is that there is more capacity in the phenomenal iconic representation than in cognition and thus that phenomenology “overflows” cognition.¹

The subjects report seeing all or almost all of the items and the cuing experiments—showing as they do “partial report superiority”—appear to back up what the subjects say. However, as Fink notes, my opponents² criticize my appeal to what the subjects say about their experience (Byrne et al. 2007). What is in the subjects’ consciousness might be just a generic representation—e.g., indicating that there is a circle of rectangles or array of letters without indicating the specific orientations of the rectangles or specific shapes of the letters. After all, we can’t expect naïve subjects to have a grip on the distinction between generic and specific phenomenology. Subjects say they have an image of all or almost all the items because they have a solely generic representation i.e., a representation that specifies the location of the items and their abstract category (letters, rectangles) but without the specific details (letter identity, orientation).

How do my opponents explain the fact that the subjects can get 7 of 8 rectangles right and 3 to 4 letters from any cued row if their phenomenal icons do not contain the specific information needed to do these tasks? According

¹ For experiments from the Lamme lab, see (Sligte et al. 2008, 2010, 2011; Vandenbroucke et al. 2011). This result has been replicated by other labs. See for example, (Freeman & Pelli 2007). My discussions of these experiments appear in (Block 2007a, 2007b, 2008). See also (Jacobson 2014) for a discussion of a different relationship between the dissociation between access and phenomenal consciousness and the dissociation between phenomenal character and representational content.

² He references (Grush 2007) but the point is also made in other critiques (Kouider et al. 2007; Papineau 2007; Van Gulick 2007)

to these opponents, the specific details of the shapes are registered unconsciously. And when subjects think they are reading details off of an already present conscious image, what they are really doing is making unconscious details conscious (Phillips 2011). Fink concludes that “Allowing SGP thus blocks Block.” (p. 10)

My response to Fink consists of three points: (1) my argument for “overflow” does not require any blanket rejection of solely generic phenomenology. (2) I have not issued any such blanket rejection and I have given qualified endorsement of some kinds of solely generic phenomenology. (3) I think there are some crucial cases—notably some spatial geometry cases in which there is reason to doubt solely generic phenomenology. I will explain these points.³

Why does Fink suppose I cannot accept generic without specific phenomenology? Part of his argument is that for an imagistic representation there cannot be generic without specific phenomenology because images are “concrete”. He says

Imagistic representation... does not allow for such indeterminacy, because images exploit the isomorphisms between concreta....Introspectively, our phenomenal experiences resemble images. If phenomenal experience represents image-like, then there cannot be SGP—and p-precision seems dangerously close to being trivial; if phenomenal experience is non-imagistic,

³ For the record, I used the generic/specific distinction in earlier papers (though not using that terminology including the one that these critics were replying to. For example, in discussing the Lamme experiment in the BBS paper to which all of these opponents were replying (Block 2007a), I said:

This supports what the subjects say, and what William James said, about the phenomenology involved in this kind of case. What is both phenomenal and accessible is that there is a circle of rectangles. What is phenomenal but in a sense not accessible, is all the specific shapes of the rectangles. (p. 488)

The phenomenology as of a circle of rectangles is generic phenomenology; the phenomenology as of the specific shapes is specific phenomenology. Further, in an earlier version of the argument based on the Sperling experiment in 1995 I also appealed to a version of the generic/specific distinction, although somewhat less explicitly (Block 1995, p. 244)

Here is the description I *think* is right and that I need for my case: I am P- conscious of all (or almost all - I will omit this qualification) the letters at once, that is, jointly, and *not just as blurry or vague letters*, but as specific letters (or at least specific shapes), but I don't have access to all of them jointly, all at once. [italics added]

then we can allow for SGP and render p-precision non-trivial—but this is in tension with some of Block's other work and our introspective evidence. (p. 10)

A similar argument to his was made by Robert Van Gulick (2007) and in a different form by Rick Grush. Van Gulick says

If one holds a “movie screen of the mind” model of phenomenal consciousness, it may seem impossible that there could be letters that are phenomenally present as letters without being present as specific letter shapes. But such a model is at best problematic, and if one rejects it, then there seems no reason why the characters of which the subjects are aware could not be indeterminate in ways that exactly match their limited cognitive access to those features. (p. 529)

In my 2007 reply to Van Gulick I rejected this argument and—contrary to what Fink says about my argument—I endorsed a version of the SGP. I said

Van Gulick notes that the “movie screen of the mind” view would say that you cannot have generic phenomenology without specific phenomenology, implicitly suggesting that I am relying on the “movie screen of the mind” view, and on the fact of generic phenomenology, to argue for specific phenomenology...I reject the principle – applied by ... Van Gulick – that pictorial representation has to specify the relevant details. I call this principle the “photographic fallacy” (Block 1983). More specifically, the photographic fallacy supposes that pictorial representations have to represent details of anything in view in the manner of a prototypical photograph. To see the fallacy, note that an impressionist painter might represent a hand in broad brush strokes that do not explicitly represent the number of fingers or whether one of them has a ring. (Block 2007b, p. 533)

It may be said that endorsing generic without specific phenomenology on my part is just incoherent since it undermines my own position. Recall that the reason Fink says I cannot endorse generic without specific phenomenology is that my opponents use it to argue that what is in consciousness in the Sperling and Lamme experiments is solely generic, the specific details being perceived unconsciously. My approach has not been to issue a blanket denial of the possibility of solely generic phenomenology but rather to argue against the claim that the highly specific representations in these experiments are unconscious (Block 2007a, 2011, 2014b).⁴

Is solely generic phenomenology possible? There certainly are some intuitively plausible (though not compelling) cases. For example, if one sees a red thing in the distance one may perhaps see it as red without seeing it as having any specific shade of red. (See Stazicker 2011, forthcoming for defenses of solely generic phenomenology.) However, even if there is generic phenomenology, I think it is doubtful in some cases, notably certain spatial cases. In particular, I doubt that there can be generic phenomenology of an oriented rectangle that does not specify the rough orientation of the rectangle.⁵

My rationale for this view is partly introspective and partly a result of informal reports of imagery experiments from Stephen Kosslyn. I have discussed doing experiments on this issue with Kosslyn and Dan Reisberg.

Imagine that you are in a house, going down the stairs and out the front door. In front of you is a picket fence with a gate. You go out through the gate and walk to the corner where you mail a letter.

⁴ If you want to get a brief taste of the kind of argument I have in mind, look at one of: (Block 2014a, 2014b). In one of the articles cited (Bronfman et al. 2014), evidence is provided of specific information about uncued rows in a Sperling-like experiment. What I especially like about this experiment is that the authors provide 3 different tests of the claim that the specific information in the uncued rows is conscious.

⁵ In (Block 2011), I said “...generic conscious representations of non-square rectangles that do not specify between horizontal and vertical orientations is difficult to accept.” Note that this is not a blanket denial of the possibility of solely generic phenomenology but rather a denial of one specific kind of solely generic phenomenology. Hilla Jacobson and Hilary Putnam relate this kind of point about imagery to a principle of “cohesiveness” of the various aspects of an image (Jacobson & Putnam forthcoming).

Stop now and answer the question: which way did you turn when you went out through the gate? Kosslyn reports in conversation that when he gives such spatial vignettes to subjects they do not report that there was no particular direction. The experimental challenge is to design an experiment that distinguishes between an answer made up on the fly and an answer based on what was “already there” in the image.

To summarize so far: Fink says “Allowing SGP thus blocks Block.” I reply that my argument for “overflow” does not require any blanket rejection of solely generic phenomenology; that I have not issued any such blanket rejection; that I have endorsed the possibility of solely generic phenomenology; and that I think there are some specific cases in which solely generic phenomenology is not very plausible.

3 Is the concept of phenomenal precision incoherent?

According to Fink, if there is no solely generic phenomenology (i.e., generic without specific phenomenology) then the concept of phenomenal precision is threatened by incoherence. What is Fink’s argument for this conclusion? Suppose there is no solely generic phenomenology. Then, according to Fink, “...the p -precision of an experience is either contradictory, generally minimal, or generally maximal, which trivializes the notion.” (p. 9) And why is that? Because, according to Fink, if you experience the color of his Figure 2 as cayenne₆₆, then if you also experience it as red, then there will be no unique precision to the experience. For red has a much wider precision range (i.e., lower precision) than cayenne₆₆. His solution is to allow for experiencing it as red without experiencing it as any specific shade: generic without specific phenomenology.⁶

Let us approach this issue by asking what *the representationist should say by way of response to Fink’s concern that there will be no unique visual precision*. Then we can ask whether some version of that response is available to me.

⁶ Of course uniqueness does not require solely generic or solely specific phenomenology. Any sole level will do.

Recall that representationists must acknowledge phenomenal precision (assuming they acknowledge representational precision) since on their view, if the representational precision of one conscious perceptual representation is greater than the representational precision of another conscious perceptual representation, then the phenomenal precisions must follow suit. Phenomenal precision—on their view—is just the shadow of representational precision. But when we see a cayenne₆₆ object as cayenne₆₆, do we thereby also see it as red? It is often supposed that this is some sort of necessity (Confession: I once thought that). To his credit, Fink points out that this is false. He says (footnote 17):

Conceptual or nomological relations do not necessarily transfer to the realm of experiences. Imagine seeing an animal as a mouse. One does not thereby see it as an owner of a heart, or as a member of the phylum chordata even though all mice belong to each category necessarily.⁷

Certainly Fink is right that seeing something as a mouse does not require seeing it as a chordate. However, he thinks any experience of cayenne₆₆ is “likely” to be an experience of red. He doesn’t say how he knows this.

Here is a tempting but wrong view that I believe may stand behind what Fink says (and is also exemplified I believe in Begby 2011 and in a more complex form in Siegel 2010). Look at the cayenne₆₆ patch in Fink’s Figure 2. I know what a red thing looks like and I can tell from looking that it is red because...well...it looks red. So I visually represent it as red. Similarly, it looks colored. And a baseball bat looks like a baseball bat, so I visually represent it as a baseball bat.

However, I also know what a 1969 Chevrolet Camaro looks like, as well as what a 1961 Jaguar E-type looks like. Do I thereby visually represent the property of being a 1969 Camaro or a 1961 E-type? I know what my wife looks

like. Do I thereby have a singular visual representation that represents her? Perhaps what I am really visually representing in each of these cases is just constellations of low level properties that are recognitionally equivalent to the property of being a 1961 Jaguar E-type or to the singular property of being my wife.

I have argued that the extent of seeing-as in the sense of visual representation is not a matter for the armchair (Block 2014c). From the armchair one does not know whether something’s looking like a 1961 Jaguar E-type is a matter of representation of constellations of colors, shapes, textures, illumination, motion and other low level properties as opposed to an actual representation of the property of being a 1961 Jaguar E-type.

For example, I give evidence that we can visually represent facial expressions (high level property) and in addition constellations of colors, shapes, textures, etc. (low-level properties). The evidence is that there are distinct “adaptation” effects for both the low and high-level properties. (Adaptation is the neural “fatigue” effect underlying afterimages.) For example, if you vary the low level properties but keep the face identity (or expression or just faceness) constant, you get smaller adaptation effects, showing an extent of low level perception. And the fact that there is a residual face adaptation effect is one of many items of evidence favoring face-specific perception.

You can experience such an adaptation effect for yourself. Stare at the picture on the right for 1 minute, covering the two pictures on the left with something. Then very briefly look at the center picture asking yourself whether it looks more fearful or more angry. Now cover the two pictures on the right and stare at the picture on the left for one minute. Now look at the center picture very briefly again. It will appear to have a different expression. The center picture is a morph of a fearful face and an angry face. When you adapt to the fearful expression you are more likely to see the morph as angry-looking and conversely for adapting to the angry expression. This doesn’t prove that there is an adaptation effect for facial expression over and above adaptation effects for constellations

⁷ By “owner of a heart” he must mean some sort of biological classification (on a par with chordate) since obviously any individual mouse could lose its heart (even briefly staying alive) and still be a mouse.

of low level properties. The best one can do is form hypotheses about what those low level properties might be and vary those properties keeping expressions constant.

In addition, one can look for other signs of visual representation of faces or facial expressions. For example, faces show “visual popout”. Since typically “conjunctive” properties do not show visual popout, that fact suggests that visual representations of faces are not “conjunctive” properties and hence not conjunctions of low level features. The upshot of this and other work I cannot describe here (Block 2014c) is that it is very likely that there are representations of face-attributes such as facial expressions in addition to representations of low level properties.⁸

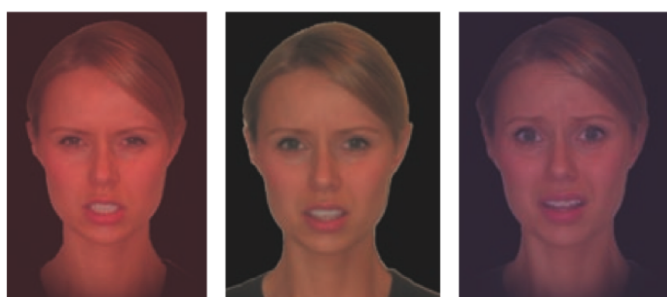


Figure 1: From Butler et al. (2008) with permission of Elsevier

The upshot of all this is that a single visual experience can represent both low level properties and high level properties. So: there can be distinct precisions for the different representations. For example, the precision of the experiential representation of fearfulness could be ascertained by investigating how much variation in the percentage of fearfulness in a morph like the middle one in the figure above is compatible with exactly the same visual representation of fearfulness. And similar methods could be used to ascertain precisions for the low level properties that are represented. There is no reason to expect these precisions to be the same.

An experience that represents cayenne₆₆ could also represent red and there could be dis-

tinct precisions for each of these representations. And what goes for representational precision also works for phenomenal precision. If more than one property is genuinely present in phenomenology then there can be distinct precisions for the distinct properties. So the solution for the representationist works even if representationism is false.

So why is there supposed to be a problem concerning unique precisions? Fink argues as follows

You might think that this color experience has two p-precision values: The first value is for being experienced as red, and the second for being experienced as cayenne₆₆. But this seems contradictory: why should one and the same experience of a color have two p-precision values, but only one for r-precision? And for that matter, why not three values for p-precision? You likely experience the color not only as cayenne₆₆ and as red, but also *as a color*? Why not four, then, if you experience it *as a visual experience*? Or five, if you experience it *as phenomenal*? [NB: p-precision is phenomenal precision; r-precision is representational precision]

The argument is not spelled out but one can guess that it depends on the idea that there is incoherence because there is no end to the number of properties that are present in experience. (Fink seems to suppose that there are not multiple representational precisions but does not say why.) We don’t need to see exactly what the argument is supposed to be to see that this premise is wrong. There is absolutely no evidence that experiences of colors present (or represent) colors as colors or as something or as phenomenal. These presentations and representations cannot be simply postulated. The reason that I went through the example of fearfulness was to give the reader a sense of how much work has to be done to show representation of a high level property. The problem in Fink’s argument is the assumption that you “likely” experience his Figure 2 not only as cayenne₆₆ but as a

⁸ In his reply to me (2014), Burge is more skeptical than I am about the power of appeals to adaptation, arguing that adaptation needs to be combined with other methods.

color and the insinuation—not explicitly stated—that you experience it as something and as phenomenal. There is simply no reason to believe this.

On my view, color experience—like all perceptual experience—is non-conceptual. But the point is even stronger if color experience is conceptual since then the concept of color and the concept of something would be required to see the cayenne₆₆ patch as colored and as something. Ask yourself whether an animal that can visually represent the color patch in Figure 2 as cayenne₆₆ must also represent it as red or as colored. Must the animal be able to attend to or notice the redness or the coloredness as well as the specific shade? Or consider 4 month old human babies whose color perception is known to be good but who do not appear to notice colors to the extent of being able to use color information to judge whether there is one or two items. Even two year old children are so bad at conceptualizing color that a term was coined in the early 20th Century, “*farbendummeit*” (color stupidity), to describe their cluelessness. Darwin thought his own children were color-blind because they were so poor at learning color names (Bornstein 1985; Campbell 2014).

To conclude this section: uniqueness of precision is not required for coherence. The representationist can reasonably hold that to the extent that there is more than one representational content, there is more than one precision: precision of representation depends on what representation is in question. And the same can be said of what properties are *presented* in perception as opposed to represented in perception, even if as I argue, representationism is false.

4 How not to clarify phenomenal precision

Here is a tempting idea about representational precision. Representational precision is just a matter of how much the stimulus can change without changing the representational content of the subsequent perception. And the same idea extends to phenomenal precision: phenomenal precision is a matter of how much the stimulus can change without changing the phenom-

enal character of the perceptual state. Of course these ideas would not be useful if one included stimulus changes that don’t make a difference when the subjects’ eyes are closed or in the dark or in a dust storm. So the proposal does not get even to first base without specifying that the circumstances of perception must be ideal.

Here is an example: Suppose one is looking at an oriented line. If a change of up to but not beyond plus or minus 1 degree makes no difference in the percept of the orientation in ideal conditions, then the representational precision is plus or minus 1 degree. And the same thought also covers phenomenal precision. If a change of up to but not beyond plus or minus 1 degree makes no difference in the phenomenology of the percept of the orientation in ideal conditions, then the phenomenal precision is plus or minus 1 degree. One advantage of this conception of precision is that representational and phenomenal precisions will be comparable. And representationist ideas can be tested. If phenomenal precisions are smaller, i.e., more precise than representational precisions, then representationism is definitely over.

I like this idea of precision for cases in which it is fairly clear what ideal conditions would consist in. But if one is concerned that phenomenal precision is not a coherent notion, this suggestion will not be of much help. The problem with this suggestion is that the notion of ideal conditions will inevitably smuggle in the ideas that are supposedly being explained. In the case of representational content, the problem has often been called the “problem of error” (Fodor 1987): representational states correlate best—not with their truth conditions—but with conditions that include systematic error. A notion of ideal conditions that avoided this consequence would itself have to distinguish between veridical and falsidical representations (see Adams & Aizawa 2010).

Fink’s proposal about phenomenal precision sometimes sounds like the correlational idea just mentioned—that phenomenal precision is a matter of how much the stimulus can change without changing the phenomenal character of the subsequent perception in ideal cir-

cumstances. However, Fink goes on to explicate the notion of *change in the phenomenology of the percept* in terms of discernability: “for example, the pain caused by 480mc/sec/cm² is not reliably discernible from one caused by 640 mc/sec/cm².” And he goes on to spell this out in terms of the lower bound on p-precision being the range of cases “one cannot distinguish by experiencing as *F* under ideal conditions.” (p. 8). In the conclusion of the paper, Fink describes his proposal in terms of the notion of a “just noticeable difference [JND] as a lower bound of p-precision.” (p. 12)

However, what one can distinguish from what is a matter not just of phenomenology but of an interaction between phenomenology and cognition. As I noted (Block 2015, sections 6 & 10), discriminability is neither necessary nor sufficient for phenomenal difference. It is not sufficient because there are sometimes ways of discriminating between percepts that do not depend on a phenomenological difference, such as beats on vibrating strings. And it is not necessary because not all phenomenological differences need be accessible to the cognitive apparatus of the subject. I mentioned phenomenal Sorites cases (Morrison 2015) in connection with this point. As has often been noted, colors A and B may be indistinguishable because the difference between color A and color B is below the JND. And B may be indistinguishable from C for the same reason even though A is distinguishable from C. One way of thinking about this is that A and B may actually look different—i.e., produce percepts with different phenomenologies, but the difference in phenomenologies may be cognitively inaccessible. If so, noticeable differences will not track phenomenal differences.⁹

In short: phenomenal precision can be explicated in terms of the extent to which the stimulus can change in ideal conditions without changing the phenomenology of the resulting percept; but explaining changes in the phenomenology of the percept in terms of noticing or in terms of discrimination brings in an interaction with cognition that ruins the explication.

I welcome Fink’s suggestions about how to explicate phenomenal precision so long as the notions of discrimination and noticing are stripped from the explication and it is acknowledged that we have no reductive account of ideal conditions. And I acknowledge the possibility of solely generic phenomenology but I don’t think it creates the problem Fink mentions for my overflow arguments.

⁹ Fink seems to acknowledge such points in footnotes 14 and 22 but somehow ignores them in explicating phenomenal precision.

References

- Adams, F. & Aizawa, K. (2010). Causal theories of mental content. *The Stanford encyclopedia of philosophy*. Stanford, CA.
- Begby, E. (2011). Review of origins of objectivity. *Notre Dame Philosophical Reviews*
- Block, N. (1983). The photographic fallacy in the debate about mental imagery. *Noûs*, 17 (4), 651-662.
- (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18 (2), 227-247. [10.1017/S0140525X00038188](https://doi.org/10.1017/S0140525X00038188)
- (2007a). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences*, 30 (5-6), 481-548. [10.1017/S0140525X07002786](https://doi.org/10.1017/S0140525X07002786)
- (2007b). Overflow, access and attention. *Behavioral and Brain Sciences*, 30 (5-6), 530-542. [10.1017/S0140525X07003111](https://doi.org/10.1017/S0140525X07003111)
- (2008). Consciousness and cognitive access. *Proceedings of the Aristotelian Society, CVIII* (3), 289-317. [10.1016/j.neures.2009.09.1651](https://doi.org/10.1016/j.neures.2009.09.1651)
- (2011). Perceptual consciousness overflows cognitive access. *Trends in Cognitive Sciences*, 15 (12), 567-575. [10.1016/j.tics.2011.11.001](https://doi.org/10.1016/j.tics.2011.11.001)
- (2014a). Consciousness, big science and conceptual clarity. In G. Marcus & J. Freeman (Eds.) *The future of the brain: Essays by the world's leading neuroscientists* (pp. 161-176). Princeton, NJ: Princeton University Press.
- (2014b). Rich conscious perception outside focal attention. *Trends in Cognitive Sciences*, 18 (9), 445-447. [10.1016/j.tics.2014.05.007](https://doi.org/10.1016/j.tics.2014.05.007)
- (2014c). Seeing-as in the light of vision science. *Philosophy and Phenomenological Research*, 89 (3), 560-573. [10.1111/phpr.12135](https://doi.org/10.1111/phpr.12135)
- (2015). Precision, acuity, veridicality and the nature of perception. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1-52). Frankfurt a. M., GER: MIND Group.
- Bornstein, M. (1985). On the development of color naming in young children: Data and theory. *Brain and Language*, 26 (1), 72-93. [10.1016/0093-934X\(85\)90029-X](https://doi.org/10.1016/0093-934X(85)90029-X)
- Bronfman, Z., Brezis, N., Jacobson, H. & Usher, M. (2014). We see more than we can report: "Cost free" color phenomenality outside focal attention. *Psychological Science*. [10.1177/0956797614532656](https://doi.org/10.1177/0956797614532656)
- Burge, T. (2014). Reply to Block: Adaptation and the upper border of perception. *Philosophy and Phenomenological Research*, 89 (3), 573-583. [10.1111/phpr.12136](https://doi.org/10.1111/phpr.12136)
- Butler, A., Oruc, I., Fox, C. J. & Barton, J. J. S. (2008). Factors contributing to the adaptation aftereffects of facial expression. *Brain Research*, 1191, 116-126. [10.1016/j.brainres.2007.10.101](https://doi.org/10.1016/j.brainres.2007.10.101)
- Byrne, A., Hilbert, D. R. & Siegel, S. (2007). Do we see more than we can access? *Behavioral and Brain Sciences*, 30 (5-6), 501-502. [10.1017/S0140525X07002816](https://doi.org/10.1017/S0140525X07002816)
- Campbell, J. (2014). Experiencing objects as mind-independent. In J. Campbell & Q. Cassam (Eds.) *Berkeley's puzzle: What does experience teach us?* (pp. 50-74). Oxford, UK: Oxford University Press.
- Fink, S. B. (2015). Phenomenal precision and some possible pitfalls: A commentary on Ned Block. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1-14). Frankfurt a. M., GER: MIND Group.
- Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. Cambridge, MA: MIT Press.
- Freeman, J. & Pelli, D. (2007). An escape from crowding. *Journal of Vision*, 7 (2), 1-14. [10.1167/7.2.22](https://doi.org/10.1167/7.2.22)
- Grush, R. (2007). A plug for generic phenomenology. *Behavioral and Brain Sciences*, 30 (5-6), 504-505. [10.1017/S0140525X07002841](https://doi.org/10.1017/S0140525X07002841)
- Jacobson, H. (2014). Phenomenal consciousness, representational content and cognitive access: a missing link between two debates. *Phenomenology and Cognitive Science*, 1-15. [10.1007/s11097-014-9399-2](https://doi.org/10.1007/s11097-014-9399-2)
- Jacobson, H. & Putnam, H. (forthcoming). Against perceptual conceptualism. *International Journal of Philosophical Studies*
- Kouider, S., de Gardelle, V. & Dupoux, E. (2007). Partial awareness and the illusion of phenomenal consciousness. *Behavioral and Brain Sciences*, 30 (5-6), 510-511. [10.1017/S0140525X07002919](https://doi.org/10.1017/S0140525X07002919)
- Lamme, V. (2003). Why visual attention and awareness are different. *Trends in Cognitive Sciences*, 7 (1), 12-18. [10.1016/S1364-6613\(02\)00013-X](https://doi.org/10.1016/S1364-6613(02)00013-X)
- Morrison, J. (2015). Anti-atomism about color representation. *Noûs*, 49 (1), 94-122. [10.1111/nous.12018](https://doi.org/10.1111/nous.12018)
- Papineau, D. (2007). Reuniting (scene) phenomenology with (scene) access. *Behavioral and Brain Sciences*, 30 (5-6), 521-521. [10.1017/S0140525X07003019](https://doi.org/10.1017/S0140525X07003019)
- Phillips, I. B. (2011). Perception and iconic memory: What Sperling doesn't show. *Mind & Language*, 26 (4), 381-411. [10.1111/j.1468-0017.2011.01422.x](https://doi.org/10.1111/j.1468-0017.2011.01422.x)
- Siegel, S. (2010). *The contents of visual experience*. Oxford, UK: Oxford University Press.
- Sligte, I. G., Scholte, H. S. & Lamme, V. (2008). Are there multiple visual short term memory stores? *Plos One*, 3 (2), e1699-e1699. [10.1371/journal.pone.0001699](https://doi.org/10.1371/journal.pone.0001699)

- Sligte, I. G., Vandenbroucke, A. R. E., Scholte, H. S. & Lamme, V. (2010). Detailed sensory memory, sloppy working memory. *Frontiers in Psychology*, 1, 1-10. [10.3389/fpsyg.2010.00175](https://doi.org/10.3389/fpsyg.2010.00175)
- Sligte, I. G., Wokke, M. E., Tesselaar, J. P., Scholte, H. S. & Lamme, V. (2011). Magnetic stimulation of the dorsolateral prefrontal cortex dissociates fragile visual short term memory from visual working memory. *Neuropsychologia*, 49 (6), 1578-1588. [10.1016/j.neuropsychologia.2010.12.010](https://doi.org/10.1016/j.neuropsychologia.2010.12.010)
- Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs*, 74 (498), 1-29.
- Stazicker, J. (2011). Attention, visual consciousness and indeterminacy. *Mind & Language*, 26 (2), 156-184. [10.1111/j.1468-0017.2011.01414.x](https://doi.org/10.1111/j.1468-0017.2011.01414.x)
- (forthcoming). The visual presence of determinable properties. In F. Dorsch, F. Macpherson & M. Nida-Rümelin (Eds.) *Phenomenal Presence*. Oxford, UK: Oxford University Press.
- Vandenbroucke, A. R. E., Sligte, I. G. & Lamme, V. (2011). Manipulations of attention dissociate fragile visual short term memory from visual working memory. *Neuropsychologia*, 49 (6), 1559-1568. [10.1016/j.neuropsychologia.2010.12.044](https://doi.org/10.1016/j.neuropsychologia.2010.12.044)
- Van Gulick, R. (2007). What if phenomenal consciousness admits of degrees? *Behavioral and Brain Sciences*, 30 (5-6), 528-529. [10.1017/S0140525X07003093](https://doi.org/10.1017/S0140525X07003093)

Rules: The Basis of Morality... ?

Paul M. Churchland

Most theories of moral knowledge, throughout history, have focused on behavior-guiding *rules*. Those theories attempt to identify which rules are the morally *valid* ones, and to identify the *source or ground* of that privileged set. The variations on this theme are many and familiar. But there is a problem here. In fact, there are several. First, many of the higher animals display a complex social order, one crucial to their biological success, and the members of such species typically display a sophisticated knowledge of what is and what is not acceptable social behavior—but those creatures have no *language* at all. They are unable even to *express* a single rule, let alone evaluate it for moral validity. Second, when we examine most other kinds of behavioral skills—playing basketball, playing the piano, playing chess—we discover that it is surpassingly *difficult* to articulate a set of discursive rules, which, if followed, would produce a skilled athlete, pianist, or chess master. And third, it would be physically impossible for a biological creature to identify *which* of its myriad rule are relevant to a given situation, and then apply them, in real time, in any case. All told, we would seem to need a new account of how our moral knowledge is stored, accessed, and applied. The present paper explores the potential, in these three regards, of recent alternative models from the computational neurosciences. The possibilities, it emerges, are considerable.

Keywords

Moral character | Moral knowledge | Moral perception | Moral rules | Neural networks | Non-discursive knowledge | Skills

Author

Paul M. Churchland
pchurchland@ucsd.edu
University of California
San Diego, CA, U.S.A.

Commentator

Hannes Boelsen
hboelsen@students.uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Editors

Thomas Metzinger
metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Introduction

An old college teacher of mine once remarked to me that “[a] philosopher’s fundamental mistakes often appear on the very first page of his major treatise”. A possible instance of this eyebrow-raising historical insight is the opening page of the long section on moral philosophy found in the prominent undergraduate philosophy textbook entitled *Introducing Philosophy*—from *Oxford University Press*, no less—skillfully edited by Robert C. Solomon (2001). Solomon there begins his broad survey of this profound and important topic with the following explanatory definition:

The core of ethics is morality. Morality is a set of fundamental rules that guide our actions.

You may well wonder how there could be anything controversial about this lucid statement, for it does indeed capture the focus of at least ninety percent of the moral philosophers’ writing in the Western traditions of religious and academic philosophy. It also captures the focus of most contemporary moral discussions, even in the marketplace and at the dinner table. We are all familiar with, and frequently argue about, presumptive “moral rules,” both major and minor. We are all familiar with the competing rationales often offered in explanation of the presumed authority of such rules—that they come from God, or that they are part of the social contract, or that (when followed) they serve to maximize collective welfare, and so forth. How *else* should we focus and pursue our con-

cern with moral reality? How else might one even *begin* to address the topic?

Hereby hangs a tale. For there are indeed other ways of approaching the topic, both as engaged citizens and as theorizing philosophers. A monomaniacal fixation on *rules* and on the source of their *authority* may reflect a fundamental misconception of what is actually going on inside successful moral agents when they engage in typical moral cognition. It may misrepresent the underlying nature of anyone's precious moral virtue. It may misrepresent the learning process by which the moral virtues are acquired. And it may misrepresent the ways in which those virtues are actually exercised in our day-to-day moral reasoning.

Before citing historical/moral authorities in hopes of winning some credence for this admittedly audacious suggestion, let us survey some of the many *non-moral*, *empirical*, or *factual* reasons for entertaining an approach to understanding morality that is not focused on rules. Such extra-moral reasons are not hard to find.

First, and perhaps foremost, rules in the literal sense require a language in which they can be expressed (and taught, and imposed, and discussed, and modified). But none of the many social creatures on this planet—excepting only humans—possess any language at all, and certainly none equal to the task of expressing even the simplest of social rules. Chimpanzees, wolves, baboons, and lions, for example, are quite innocent of language, and yet their collective behavior displays a complex social order that the adult animals must respect—on pain of punishment or retribution from their peers—and which the juveniles must learn to recognize, understand, and eventually protect with their own watchful behavior. They, too, live within a more-or-less stable moral order that serves many if not most of the same functions served by our own moral order. An adult chimp will chide, sometimes severely, a juvenile chimp that steals food from the hands of an infant chimp, and will even return the stolen food to the aggrieved victim. Wolves, and even domestic dogs, will offer comfort and solace to a wounded com-patriot and will spring to defend it against fu-

ture threats. The trust, social foresight, and mutual dependence displayed by a pack of lions organizing and executing a hunt to bring down a gazelle is a marvelous example of collective purposeful activity. And the subsequent sharing of the spoils among all who participated in the hunt is a striking example of distributive justice, even if momentary squabbles occasionally break out over access to the choicest bits of the kill. (Nobody is morally perfect, especially a tired and hungry lion.)

In sum, moral perception, moral reasoning, moral activity, moral norms, moral defense, and moral retribution all exist elsewhere in the animal kingdom (presumably for many of the same reasons that they exist in us), but in none of those other cases do language or discursive rules play any role at all in the moral phenomena at issue. The whole thing happens—most of it, anyway—but without language.

So what is going on? What is it that regulates or steers their behavior, if not rules? Before canvassing possible answers to this question, let us ponder some additional data, this time concerning humans. Adult humans occasionally fall victim to something called *global aphasia*, a stroke-induced brain malady in which the cortical areas responsible for the manipulation, production, and comprehension of *language*—in any form: spoken, written, or printed—are totally destroyed. The loss of this critically important neuronal machinery (roughly, Broca's area and Wernicke's area, typically on the left side of the brain) leaves the victim without any capacity to formulate, process, or comprehend any linguistic structures whatsoever. That dimension of cognitive representation is now completely out of business. There is nothing wrong with the victim's sensory inputs or motor outputs; these peripheral systems remain entirely functional. The cognitive deficit lies deeper. The capacity for even *forming* linguistically structured thoughts has disappeared entirely. The victim cannot formulate or comprehend any declarative sentence, nor any interrogative sentence, nor any imperative sentence, nor any rule. These elements, so familiar to the rest of us, no longer play any role in their cognitive lives.

And yet their cognitive lives in other respects remain surprisingly unaffected, despite this disaster where specifically *linguistic* structures are concerned. Some three decades ago, we had such a left-brain stroke victim in our own extended family. Aunt Betty, as she was fondly called, could still drive a car around town, shop for the groceries, cook a dinner, and watch a football game on TV with understanding and enjoyment. More to the point, her basic trust in other humans, and her own basic trustworthiness, were quite intact. During visits, her comprehension of the moral flux around her, especially where the adventures and interactions of our youngish children were concerned, seemed quite undiminished, as were her skills in providing comfort for the teary-eyed and fairness in the distribution of small pastries at lunch. Her moral cognition was up and running smoothly, evidently, much as before—but without the benefit of any rules to tell her what to do. She could no longer comprehend or even contemplate them, and yet somehow, she didn't need them.

Another illustration of the superfluity of rules to moral character emerged, without warning and to much amusement, in an interview of a moderately charming Georgia Congressman on the TV comedy show *The Colbert Report*. The topic of their extended discussion was a recent higher-court ban on the public display of the Judeo-Christian Ten Commandments in the foyer of a Louisiana courthouse, and the justice/injustice of their subsequent court-ordered removal from that public venue. The congressman, a Mr. Lynn Westmoreland, was defending their public, cast-bronze-on-granite display on a variety of grounds, but most trenchantly on the grounds that, collectively, those ten rules constitute the very foundation of our morality, insofar as we have any morality. Their public display, therefore, could only serve to enhance the level of individual morality.

Sensing an opportunity, Steven Colbert nodded his presumptive assent to this claim, and asked his guest, “Could you please cite them for us, congressman?” Westmoreland, plainly taken aback by the request, gamely began, “Don't lie, . . . don't steal, . . . don't kill,

. . .” as Colbert, with his eyebrows raised in expectation, held up first one finger in response, then two, then three. After an awkward pause at that point, the congressman, who had plainly drawn a blank beyond those three, bravely and with evident honesty said, “No, I'm sorry. I can't name them all”. My immediate reaction (oh, alright, my second) was sympathy for the congressman, because I don't think I could have named them all, either. At which point Colbert ostentatiously thanked his guest for his wisdom and brought the interview, before a large audience, to an uproariously received and laughter-filled conclusion.

The comedic point was plain enough and doesn't need any further elaboration from me. But there is a deeper lesson to be drawn from this exchange. The fact is, the congressman is probably as good an example of worthy moral character as one is likely to encounter at one's local post office or grocery store. After all, he inspired sufficient public trust to get himself elected, and he thinks morality important enough to defend it, with some passion and resourcefulness, on television. He is a presumptive example of a conscientious man with a morally worthy character. But if he is, these welcome virtues are clearly *not* owed to his carrying around, in memory, a specific list of discursive rules, rules at his immediate command, rules that he literally consults in order to guide his ongoing social behavior. He could remember only three of the ten “commandments” at issue, and, if you check the bible, he didn't get two of those three quite right in any case. If we are looking (and we *are*) for an explanation of the actual ground or source of people's moral behavior, the proposal that we are all following a specific and finite set of discursive *rules* in order to produce that behavior is starting to look strained and threadbare, to put it mildly.

Before addressing an alternative explanation, let us note one further domain of empirical evidence, relevant to our issue concerning the role of rules. Moral expertise is among the most precious of our human virtues, but it is not the only one. There are many other domains of expertise. Consider the consummate skills displayed by a concert pianist, or an all-star bas-

ketball player, or a grandmaster chess champion. In these cases, too, the specific expertise at issue is acquired only slowly, with much practice sustained over a period of years. And here also, the expertise displayed far exceeds what might possibly be captured in a set of discursive rules consciously followed, on a second-by-second basis, by the skilled individuals at issue. Such skills are deeply inarticulate in the straightforward sense that the expert who possesses them is unable to simply *tell* an aspiring novice *what to do* so as to be an expert pianist, an effective point guard, or a skilled chess player. The knowledge necessary clearly cannot be conveyed in that fashion. The skills cited are all cases of knowing *how* rather than cases of knowing *that*. Acquiring them takes a lot of time and a lot of practice.

To be sure, the point-guard can instruct the novice, “When you get possession of the ball at your end, dribble it down the floor toward the opposition’s basket, and when the defense starts to resist, pass the ball to whichever of your teammates has the best chance of sinking a shot.” But this rule, even if it is tattooed on the novice’s forearm, will hardly make him an effective player. It doesn’t tell him how to *dribble* effectively, nor could any other list of rules. It doesn’t tell him how to *recognize* a teammate’s fleeting opportunity to take a high-percentage shot, or perhaps set one up for yet a third player. It doesn’t tell him how to *pass* the ball so as to avoid interception, or how to *deceive* the defense with various kinds of fakes and feints. It doesn’t even address the issue of how to execute any one of the dozen or so different kinds of shots he himself might have to take, or when to take them. It doesn’t tell him .01 percent of what he needs to know to be a skilled player. And even if he did somehow memorize 10,000 rules on all of these diverse topics, he couldn’t possibly recall, from that vast store, exactly the rule relevant at any instant and then apply it swiftly enough to steer his ongoing play. The game unfolds much too quickly for that plodding strategy to be effective. Something else is going on inside the basketball player’s head. Something else entirely.

The game of chess is much slower, of course, and simpler too. But the same lesson emerges here as well, although from an unexpected direction. Unlike the basketball case, and because of the discreteness and comparative simplicity of chess, computer programmers have indeed written computer programs—that is, large sets of literal rules for the computer to consult and follow—that will enable a computer to play a creditable game of competitive chess. These programs were common by the early 1980s, and they were competent enough to defeat non-expert human chess players (such as me) quite regularly.

The computer-guiding rules were written so as to address any arbitrary configuration of chess pieces on the board, as might emerge in the course of a game, and to evaluate, in sequence, the cost or benefit of each of the perhaps thirty legal moves (or something in that neighborhood—it will vary) then available to the computer. To be at all effective, this strategy requires that the computer also considers the potential cost/benefit (to the computer) of its opponent’s possible *responses* to each of those contemplated moves. Each such response would of course present the computer with a new set of possible moves of its own, each requiring evaluation, and so on, for another cycle of possible moves-and-responses. If the computer is to look ahead in this fashion for only two cycles of play, it will already be evaluating something like $(30 \times 30) \times (30 \times 30) = 810,000$ or almost a million possible move-sequences! And if it presumes to look forward, in this brute-force evaluative fashion, a mere *four* cycles of play, its task explodes to examining the cost/benefit ratio for almost a *trillion* possible move-sequences.

Now you and I could never hope to execute a game-strategy of this kind, but a computer can, although just barely. Let us assume that the computer’s central processing unit (CPU) has a clock-frequency of, say, 100 Megahertz (= 100 million elementary computations per second), a fairly modest machine, these days. Such a computer will take only $(1 \text{ trillion moves to be evaluated}) / (100 \text{ million evals/sec}) = 10,000 \text{ seconds}$, or about three hours to com-

plete its evaluation of four cycles of play, assuming that the cost/benefit estimate for each move-sequence (a comparatively simple matter) can be calculated in a single elementary computation.

“But this is still ridiculous,” you might say. “Three *hours* of mulling per turn!? That’s not even legal. And looking ahead only four move-cycles? *That’s* not going to defeat a really good human chess player.” And you would be right. But in fact, some artful pruning of the decision-tree constructed by the computer’s program (e.g., through ignoring some possible moves, on both sides, that are likely to be irrelevant) will substantially reduce the combinatorial explosion in the number of moves that need to be evaluated. This can reduce the time of evaluation from three hours to perhaps three minutes, though at some cost to security. A somewhat faster CPU might further reduce it to less than three seconds. And the occasional deployment of a slightly more penetrating five or six-cycle lookahead evaluation for the occasional moves of potentially great value, positive or negative, can add some deeper, if localized, insight without adding too much in the way of a computational burden. In these ways the programmed computer can be brought into the range of real-time chess competence, even excellence.

Still, it is worth remarking that it took over three decades of program and computer development before a chess-playing computer was finally able to defeat a world-champion human chess master. The Russian master Gary Kasparov (poor devil) finally went down to an IBM monster computer named “Deep Blue” in 1997, to the celebration of nerds and technophiles everywhere (Campbell et al. 2002). That is, the gross strategy of applying discursive rules, again and again at blistering speeds, finally paid off. But it did so only because the computer CPU’s clock-speed was roughly a million times faster than any cyclic process in a human brain (which maxes out at a mere one hundred cycles per second) and only because the conduction velocity of the electrical signals inside the computer (almost the speed of light) was roughly a million times faster than the conduction velo-

city in a human nerve fiber (about the speed of a fast bicycle rider). These make the computer about (a million times a million =) a *trillion* times faster than we are. Without these singular and *superhuman* physical advantages, the computer and its list of rules—its program—would be dead in the water. And so would we humans, if the rule-based strategy were how human chess-playing competence is grounded. But plainly it is not. It couldn’t possibly be. Something else is going on inside the human chess-master’s head. Something else entirely.

2 An alternative account of moral skill

We have only recently begun to understand what that “something else” is. It has to do with the peculiar way the brain is wired up at the level of its many billions of neurons. It also has to do with the very different style of representation and computation that this peculiar pattern of connectivity makes possible. The basics are quite easily grasped, so without further ado, let us place them before you.

The first difference between a conventional digital computer and a biological brain is the way in which the brain *represents* the fleeting states of the world around it. The retinal surface at the back of your eye, for example, represents the scene currently before you with a pattern of simultaneous (repeat: simultaneous) activation or excitation levels across the entire population of rod- and cone-cells spread across that light-sensitive surface. Notice that this style of representation is entirely familiar to you. You confront an example of it every time you watch television. Your TV screen represents your nightly news anchor’s face, for example, by a specific pattern of brightness levels (“activation” levels) across the entire population of tiny pixels that make up the screen. Those pixels are always there. (Tiptoe up to the screen and take a closer look.) What changes from image to image is the *pattern* of brightness levels that those unmoving pixels collectively assume. Change the pattern and you change the image.

It is the same story with any specialized population of neurons, such as the retina in the eye, the visual cortex at the back of the brain,

the cochlea of the inner ear, the auditory cortex, the olfactory cortex, the somatosensory cortex, and so on and so on. All of these neuronal areas, and many others, are specialized for the representation of some aspect or other of the reality around us: sights, sounds, odors, tactile and motor events, even features of social reality, such as facial expressions. These neuronal activation-patterns need not be literal *pictures* of reality, as they happen to be in the special case of the eye's retinal neurons. But they are *representations* of the fine-grained structure of some aspect of reality even so, for each activation-pattern contains an enormous amount of *information* about the external feature of reality that, via the senses and internal brain pathways, ultimately produced it.

Just *how much* information is worth noting. The retina contains roughly 100 million light-sensitive rods and cones. (In modern electronic camera-speak, it has a rating of one hundred megapixels. In other words, your humble retina still has *ten times* the resolution of the best available commercial cameras.) Compare this to the paltry representational power of a typical computer's CPU: it might represent at most a mere 8 bits at a time, if it is an old model, but more likely 16 bits or 32 bits for a current machine, or perhaps 64 or 128 bits for a really high-end machine. Pitiful! Even an old-fashioned TV screen simultaneously activates about 200,000 pixels, and an HDTV will have over two-thirds of a million ($1,080 \times 640 = 691,200$ pixels). Much better. But the retina, and any other specialized population of neurons tucked away somewhere in the brain, will have roughly 100 million simultaneously activated pixels. Downright excellent. Moreover, these pixels—the individual neurons themselves—are not limited to being either on or off (i.e., to displaying a one or a zero), as with the elements in a computer's CPU. Biological pixels can display a smooth variety of different excitation levels between the extremes of 0 percent and 100 percent activation. This smooth variation (as opposed to the discrete on/off coding of a computer's bit-register) increases the information-carrying capacity of the overall population dramatically. In all, the representational technique

deployed in biological brains—called *population coding* because it uses the entire population of neurons simultaneously—is an extraordinarily effective technique.

The brain's *computational* technique, which dovetails sweetly with its representational style, is even more impressive. (As with any computer, a computational operation in the brain consists in its *transforming* some input representation into some output representation.) Recall that any given representation within the brain typically involves many millions of elements. This poses a *prima facie* problem, namely, how to deal, swiftly, with so many elements. Fortunately, what the brain *cannot* spread out over *time*—as we noted above, it is far too slow to use that strategy—it spreads out over *space*. It performs its distinct elementary computations, many trillions of them, each one at a distinct micro-place in the brain, but all of them at the *same time*. Let us explain with a picture so you can see the point at a glance.

At the bottom of figure 1 is a cartoon population of many neurons—retinal neurons, let us suppose. As you can see, they are currently representing a human face, evidently a happy one. But if the rest of the brain is to *recognize* the specific emotional state implicit in that sensory image, it must *process* the information therein contained so as to activate a specific pattern within the secondary patch of neurons just above it. That second population, let us further suppose, has the proprietary job of representing any one of a range of possible *emotions*, such as happiness, sadness, anger, fear, boredom, and so forth. The system achieves this aim by sending the entire retinal activation-pattern upward via a large number of signal-carrying axonal fibers, each one of which branches at its upper end to make fully eighty *synaptic connections* with the neurons at this second layer. (Only some of these axonal fibers are here displayed, so as to avoid an impenetrable clutter in the diagram. But every retinal neuron sends an axon upward.)

When the original retinal activation-pattern reaches the second layer of emotion-coding neurons, you can see that it is forced to go through the intervening filter of (4,096 axons \times

80 end-branches each = 327,680) almost a third of a million synapses, *all at the same time*. Each synaptic connection magnifies, or muffles, its own tiny part of the incoming retinal pattern, so as collectively to stimulate a *new* activation-pattern across the second layer of neurons. That new pattern is a representation of a specific emotion, in this case, happiness. The third and final layer of this neural network has the job of discriminating these new 80-element patterns, one from another, so as to activate a single cell that codes specifically for the emotion still opaquely represented at the second population. That is achieved by tuning a further population of synaptic connections from every cell in the middle layer to each of the five cells in the final layer. In all, what was only *implicit* in the original retinal activation-pattern (mostly in the mouth and eyebrows) is now represented *explicitly* in the top-most activation-pattern across the five cells there located.

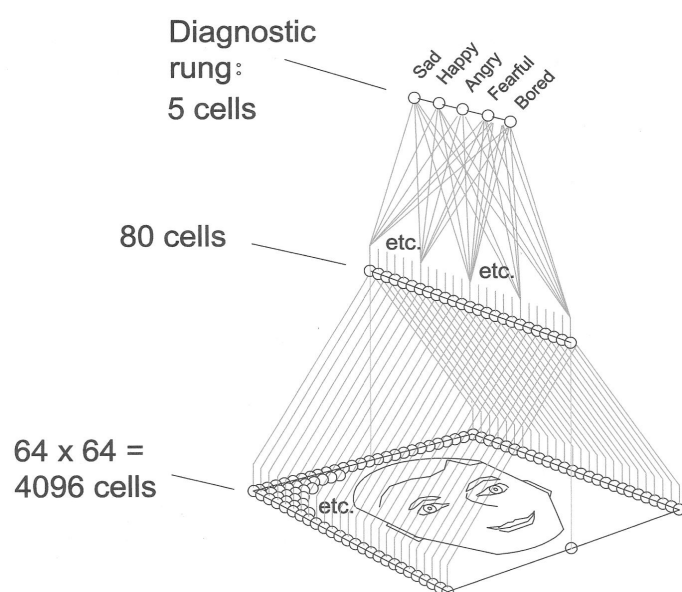


Figure 1

This trick is swiftly turned by the special configuration of the various *strengths* of each of the intervening synaptic connections. Some of them are very large and have a major impact in exciting the upper-level neuron to which it is attached, even for a fairly weak signal arriving from its retinal cell. Other connections are quite small and have very little excitatory impact on the receiving cell, even if the arriving retinal

signal is fairly strong. Collectively, those 327,680 synaptic connections have been carefully adjusted or tuned, by prior *learning*, to be maximally and selectively sensitive to just those aspects of any face image that convey information about the five emotions mentioned earlier, and to be “blind” to anything else. The complex “pattern transformation” they effect plainly loses an awful lot of information contained in the original (retinal) representation. Indeed, it loses most of it. But it does succeed in making explicit the specifically emotional information that this little three-layer “neural network” was designed to detect.

This style of computation is called Parallel Distributed Processing (PDP), and it is your brain’s principal mode of doing business on any topic. Even in this cartoon example, you can see some of the dramatic advantages it has over the “serial” processing used in a digital computer. A typical 8-bit CPU has a population of only eight representational cells at work at any given instant, compared to fully 4,096 just for the sensory layer of our little cartoon neural network. The CPU performs only eight elementary transformations at a time, as opposed to 327,680 for the neural network, one for each of its 327,680 synaptic connections. When we consider the human brain as a whole instead of the tiny cartoon network above, we are looking at a system that contains roughly a thousand distinct neuronal populations of the same size as the human retina, all of them interconnected in the same fashion as in the cartoon. This gives us (1,000 specialized populations × 100 million neurons per population =) a total of 100 *billion* neurons in the brain as a whole. As well, the total number of synaptic *connections* there reaches more than 100 *trillion*, each one of which can perform its proprietary magnification or minification of its arriving axonal message at the very same time as every other. Accordingly, the brain doesn’t have to do these elementary computations in laborious temporal sequence in the fashion of a digital computer. As we saw, a PDP network is capable of pulling out subtle and sophisticated information from a gigantic sensory representation *all in one fell swoop*. That is the take-home lesson of our cartoon net-

work. The digital/serial CPU is doomed to be a comparative dunce in that regard, however artificial may be the *rules* that make up its computer program. They simply take too long to apply.

Enough of the numbers. What wants remembering in what follows is the holistic character of the brain's representational and computational activities, a high-volume character that allows the brain to make penetrating interpretations of highly complex sensory situations in the twinkling of an eye. You are of course intimately familiar with this style of cognition: you use it all the time. Every time you recognize frustration in someone's face, evasion in someone's voice, hostility in someone's gesture, sympathy in someone's expression, or uncertainty in someone's reply, a larger version of the neuronal mechanism in figure 1 has made that subtle information almost instantly available to you.

Now, however, you know *how*: massively parallel processing in a massively parallel neural network. Or, to put it more cautiously, almost three decades of exploring the computational properties of artificial neural networks, and almost three decades of experimenting on the activities of biological neural networks have left us with the hypothesis on display above as the best hypothesis currently available for how the brain both represents and processes information about the world. No doubt, the special network processor inside you, the one that is responsible for filtering out specifically emotional information, has more than the mere two layers depicted in our cartoon. In fact, anatomical data suggests that your version of that network has the retinal information climb through four or five distinct neuronal layers before reaching the relevant layer(s), deep in the brain, that explicitly registers the emotional information at issue. The original retinal information will thus have to go through four or five distinct layers of synaptic filters/transformers before the emotional information is successfully isolated and identified. But that still gives us the capacity for recognizing emotions in less than a few tenths of a second. (The several neuronal layers involved are only ten milliseconds apart.) On matters like this, we are *fast*, at least when our myriad synaptic connections have been appropriately tuned up.

The PDP hypothesis also gives us the best available account of how that synaptic tuning takes place, that is, of how the brain *learns*. Specifically, the size or "weight" of the brain's many transforming synapses *changes* over time in response to the external patterns that it repeatedly encounters in experience. The overall configuration of those synaptic connections and their adjustable weights is gradually shaped by the recurring themes, properties, structures, behaviors, dilemmas, and rewards that the world throws at them. The resulting configuration of synaptic weights is thus made selectively sensitive to—one might indeed say *tuned* to—the important features of the typical environment in which the creature lives. In our case, that environment includes other people, and the pre-existing structure of mutual interaction and social commerce—the *moral order*—in which they live. Learning the general structure of that pre-existing social space, learning to recognize the current position of oneself and others within it, and learning to navigate that abstract space without personal or social disasters are among the most important things a normal human will ever learn.

It takes time, of course. An infant, before his first birthday, can distinguish between sadness and happiness, but little else. A grade-school child can pick up on most of the more subtle emotional flavors listed three paragraphs ago, though probably only in the behavior of young children like themselves. But a normal adult can detect all of those flavors, and more, quickly and reliably, as displayed by almost any person she may encounter. (Only psychopaths defeat us, and that's because they have deviant or truncated emotional profiles.)

Withal, learning to read emotions is only a part of the perceptual and interpretational skills that normal humans acquire. People also learn to pick up on people's background *desires* and their current practical purposes. We learn to divine people's background *beliefs* and the current palette of factual information that is (or isn't) available to them. We learn to recognize who is bright and who is dull, who is kind and who is mean, and who has real social skills and who is a fumbling jerk. Finally, we learn to *do* things. We learn how to win the trust of others, and how to

maintain it through thick and thin. We learn how to engage in cooperative endeavors and to do what others rightfully expect of us. We learn to see social trouble coming and to head it off artfully. And we learn to apologize for and to recover from our own inevitable social mistakes.

These skills of moral cognitive *output* (i.e., our moral behavior) are embodied in the same sorts of many-layered neural networks that sustain moral cognitive *input* (i.e., our moral perception). The diverse cognitive interpretations produced by our capacity for moral perceptions are swiftly and smoothly transformed—again by a sequence of well-trained synaptic filters/transformers—into patterns of excitation across our *motor* neurons (which project to and activate the body's muscles) and thereby into overt social behaviors, behaviors that are appropriate in light of the moral interpretations that produced them. Or at least, they will be appropriate if our moral education has been effective.

This weave of perceptual, cognitive, and executive skills is all rooted in, and managed by, the intricately tuned synaptic connections that intervene between hundreds of distinct neuronal populations, each of which has the job of representing some proprietary aspect of human psychological and social reality. That precious and hard-won configuration of synaptic weights literally constitutes the social and moral *wisdom* that one has managed to acquire. It embodies the unique profile of one's moral *character*: it dictates how we see the social world around us, and it dictates our every move within it. It is not an exaggeration to say that it dictates who we are. If our character needs changing or correcting, it is our myriad synapses that need to be reconfigured, at least in minor and perhaps in major ways. In all of these matters, then, don't think *rules*. Think information-transforming *configurations of synaptic weights*, for it is they that are doing the real work.

3 Reconceiving moral competence in non-classical terms

What *is* that “real work”? If the neural networks that make up our brains are not in the business of applying rules, vast libraries-full of

them, just what business *are* they engaged in? How should we think of what they are doing, if not as administering rules? What is the positive alternative to this traditional construal, expressed in non-technical language?

What those networks are doing is (trying to) interpret any *new* experience or situation as being an instance of some prior category that the brain already understands. They are trying to assimilate each new social/moral situation to an already grasped prototype situation, a template or prototype that has been incrementally created by the brain's prior experience with its surrounding social/moral reality. They are trying to grasp each of the endless novelties that they encounter as being just a modified case of some kind-of-thing that they have already encountered many times, and with which they have already become familiar. They are trying to interpret the fleeting here-and-now (which is always specific) in terms of their comparatively enduring background concepts (which are always general). They are trying to identify which of their various categories, categories that past experience has constructed for them, is the one into which their current experience fits most closely and most accurately. In sum, they are trying to apply their acquired conceptual and practical wisdom to their current situation.

Why should they, or rather, you, be trying to do that? For the very good reason that your acquired concepts or prototypes are precisely what contains your accumulated information about the world, information beyond what your current and highly specific experience happens to make evident. Those abstract prototypes contain presumptive information about the wide range of *features* that any instance of an applied concept can typically be expected to display, about the wide range of *relations* it will typically bear to other things, about the ways in which it will typically unfold or *behave* over time, and about the ways in which it can typically be *controlled* or *steered*. That is the point, after all, of having a conceptual framework in the first place. It embodies your accumulated understanding of the world's enduring background structure, your grasp of the unchanging background framework within which the ephemer-

eral and the changeable are always constrained to unfold.

Consider, for an example of moral perception in particular, the arrival of lunchtime in a typical elementary-school classroom. Every student retrieves a paper-bag lunch from the cloak-room and settles down to consume its contents. You are one of those students and, while eating, you perceive Johnny surreptitiously attempt to remove a banana from the lunch-bag next to Michael. On the face of it, you are witnessing a case of theft. And that interpretation implies many things: that the banana belongs to Michael, that Michael will be seriously aggrieved when he discovers Johnny's affront, that Johnny has inadequate self-control, that a noisy conflict will ensue if events are left to themselves, and so on and so on. This situation, as described, warrants some immediate intervention.

Most obviously, you might just openly berate Johnny in front of the other students. Or, more boldly, you might seize the banana from Johnny and quietly return it to Michael. Or you might call the teacher and rat Johnny out. These hardly exhaust your possible responses, but they are all typical sorts of responses to a typical sort of problem, and which response you choose will depend on contextual factors such as how big and mean Johnny is, how susceptible he is to collective disapproval, and how reliable the teacher is at dispensing justice. Perhaps the first path is the best response, with the second and third left as backups if the first path fails to return the situation to a just equilibrium.

And so that is what you do: berate him on the spot. All within a second of witnessing the presumed theft. Because your eight-year-old brain is already keenly tuned to that sort of possibility and to thousands of other social possibilities as well. Given your well-trained neural networks, it takes only the external perceptual situation itself to provoke the interpretation of theft. And it takes only that conceptual interpretation itself, in the context of one's ever-present character and background information, to activate your overt social response.

Your interpretation, of course, might be incorrect. Perhaps Johnny was just trying to retrieve his own banana, earlier stolen by the avari-

cious Michael. Perhaps your openly berating *Johnny* was inappropriate, since everyone in the class except you witnessed Michael's earlier theft but was too frightened of Michael to do anything about it. If so, Johnny has now been victimized twice over, once by Michael and once by you.

To be sure, there are many other convoluted possibilities, in addition to or beyond this one. But they are increasingly unlikely, compared to your first take on the situation. This is why your brain fell so swiftly and easily into that straightforward interpretation: *theft* is the simplest, most obvious, most probable explanation of what you have actually seen, and that's why it's the explanation that the brain tries first. Furthermore, once that explanatory assumption is in place, an immediate attempt at restitution is the most natural expression of your antecedent character and your acquired social skills.

What is impressive here is not just the swiftness with which your cognitive resources get tapped. It is the enormous *range* of alternative possibilities to which your brain is/was no less prepared to respond, and with equal swiftness, insight, and know-how. If, instead of a banana theft, you had witnessed Mary accidentally press her hand against the point of a newly sharpened pencil, your recognition of her pain and your comforting response would have been just as quick. If you saw the class's pet rabbit escape from its (poorly locked) cage, you would know to retrieve it and return it to its proper home. If you had turned to see a small fire blazing in the classroom's bookcase-corner library, with Johnny (him again!) slipping a plastic lighter into his pocket, you would grasp the significance of the event instantly and let out a loud warning to everyone in the room. If (here we deliberately choose something unlikely) Superman, with cape swirling, then bursts through the open classroom window and asks, "Which way did the fire-bug go?!", you would know to point to Johnny's fleeing backside as he hightails it out the classroom door. If . . . if . . . if . . . for a thousand thousand "ifs" and more, even your eight-year-old self would be competent to recognize the situation and to respond to it swiftly and appropriately.

This extraordinary breadth of capacity is a consequence, in part, of the *combinatorics* of

the already large number of neurons the brain uses to represent any sort of social situation. It is the same trick, once again, used by your television screen, in order to display an almost endless variety of possible pictures, despite a (large but) finite set of pixels with which to portray them. The retina of the eye uses the same trick, recall, but boasts many more “pixels” than a TV screen. Your perceptual capacities, accordingly, far exceed the modest range of that familiar technology.

Of course, simply representing something at the perceptual level does not mean that you *understand* it, and that is strictly what concerns us here. To understand a perceptual input is, as we saw above, to assimilate it to one of the brain’s learned *prototype* situations, to one of the standard, recurring, well-patterned kind of circumstances that one’s past experience has impressed upon your memory and your habits of behavior. That memory and those habits, you will recall, are a matter of the acquired configuration of the brain’s synaptic connections and their synaptic strengths or “weights.” For it is those collective synaptic “filters” or “transformers”—the ones that intervene between each of the brain’s many reality-portraying neuron populations—that steer the initial perceptual representations into the higher-level prototype patterns that fit those percepts most closely.

Look again at the toy network of figure 1 and note that its 327,680 synaptic connections were there adequate to steer a wide variety of possible input face images into one or other of exactly five emotional prototypes. If we suppose that this ratio (i.e., 327,680 synapses for every five categorial prototypes) is roughly typical, then a brain like yours, with 100 *trillion* synaptic connections, should be able to learn, and to deploy immediately (when appropriate), something in the neighborhood of $(5 / 327,680) \times 100 \text{ trillion} = 1.5 \text{ billion}$ distinct categorial prototypes! Now, presumably we don’t have quite *that* many distinct categories awaiting activation. That number is best viewed as a theoretical upper limit on what we might achieve. But in light of how our cognitive systems evidently do their jobs, it is small wonder that even your grade-school self is hair-trigger ready for

such an astonishing range of situations, social and otherwise—and ready, note well, with an astonishing range of understanding and relevant skills.

4 Moral conflict and moral reasoning

Alas, our cognitive systems don’t always work perfectly. Sometimes we misinterpret what we are seeing and hearing. That is, sometimes we assimilate the case at hand to a category or prototype of which it is not an instance, to which it positively does not belong. When that happens, you become the victim of the entire family of expected features, relations, developmental profile, and presumptively appropriate behavioral responses that automatically come with that prototype, but that fail to accurately characterize or suit the case at hand. Some dimensions of the activated prototype may fit (that’s why you deployed it in the first place), but others do not, as you slowly come to appreciate. As the case before you unfolds, and perhaps as you learn more about its initial stages, your prototype-driven expectations are violated and your cognitive dissonance grows. You have somehow failed to understand the situation correctly.

At some point, the accumulated new input or evidence may be sufficient to kick your brain’s activational activity *out* of the prototype-category that initially captured it and *into* a different and more appropriate prototype, one whose overall profile finally does fit the case at hand. At that point you may have the familiar “click” experience, where the problematic case suddenly re-presents itself in a new and coherent light, and you think to yourself, “Oh my god, I misunderstood what was happening.” You may then struggle to repair the social/moral damage that your automatic but ultimately inapt behavior may have produced.

This happens to all of us, and quite often. It reflects the fact that our moral cognition is not infallible. Happily, such mistakes can be corrected, and regularly they are, sometimes by oneself and sometimes with the help of others. Unhappily, sometimes they are not corrected. We are all familiar with people who have too

quickly taken a superficial interpretation of some social/moral issue and then stubbornly refuse to respond to, or even to see, its failures to capture adequately the social/moral complexities that the issue presents.

When this happens, we have a typical case of moral conflict. If the issue is pressing, we may begin a round of moral reasoning and moral argument with the person or persons who take the competing interpretation of the issue, and who propose a problematic response or policy in light of it. Such arguments, it must be admitted, often begin with both sides citing some favored “moral” or other, a rule that supposedly compels us to take their response to the situation or to embrace their policy recommendation. But this rarely settles the conflict, since the real disagreement is usually about how we should *interpret* the situation in the first place.

Classic examples are right in front of us. The public debate over abortion involves a presumptive conflict between the rule “Any innocent human person has the moral right to continue living” and the rule “Any woman has the moral right to control her own internal reproductive activities.” But the debates typically focus on how these rules should be *interpreted*, what *qualifications*, if any, should limit their application, and which of these conflicting rules carries the greater *authority*. Ultimately, as both sides of the debate usually *agree*, the issue boils down to whether or not the fertilized egg and/or the early fetus that develops therefrom really is, or should be counted as, a *human person* in the first place. The right-to-life folks say “yes.” The defenders of choice say “no.”

Our point in rehearsing this issue is that, even in the case of this most celebrated of moral conflicts, the primary issue, once again, is not really about rules. It is about how we should interpret or categorize, rationally and accurately, the early fetus. One side will argue, “It’s just a clutch of unfolding stem cells, without a brain or nervous system, without any character or personal identity, without any will or consciousness, without any of the dimensions of genuine personhood. It is no more a person than a recently-planted acorn is already an oak tree.” The other side will argue, “Personhood begins

at conception, at fertilization. That is when God places a human soul into the now-developing egg. Accordingly, that is when the right to life begins, a right not to be subsequently denied. (And by the way, acorns don’t have immaterial souls.)”

The first side will respond, “We don’t accept your utterly unverified theory of immaterial souls implanted by a divine being at conception, and we resist your attempt to thus impose your arbitrary and fantastical religious beliefs on the rest of us. (And by the way, modern science implies that humans don’t have immaterial souls either.)” To which the second side will counter, “Your position acknowledges *no* clear or well-defined point at which the developing fetus begins to acquire rights. If it is acceptable to terminate the life of the developing fetus, why isn’t it acceptable to terminate the life of a developing *newborn baby*? That would plainly be over the top, but the case of a fetus is different in no fundamental respect.”

And so it goes. Each side of the debate typically attempts to get the other side to see the problematic case “in a different light,” to interpret it as relevantly similar to a distinct but salient prototype whose moral status is not under dispute, to assimilate it to a category that is factually more adequate to the problematic case at hand. Thus the category “mindless clutch of cells” vies with the category “innocent and defenseless person” for our cognitive apprehension of the conceptus and early fetus. Arguments here are not conducted by repeatedly citing moral rules and deducing consequences therefrom. They are conducted by repeated attempts to highlight diverse factual similarities, and dissimilarities, between each of the contesting moral prototypes, on the one hand, and the conceptus/early fetus on the other.

I deploy this example of a moral disagreement and its typical discussion not to try to settle the issue in favor of either side here, but to illustrate the forms that moral disagreements and moral arguments typically display. It is, most assuredly, *not* the aim of this naturalistic and brain-focused essay to try to deduce any substantive moral rules from our growing understanding of how the brain conducts its moral

cognition. Brains arrive at their moral wisdom by a long process of learning, often painful learning, whether in the lifetime of an individual or in the centuries-long development of a society, and there is no substitute for this learning process. It is rather like the development of *scientific* wisdom, if I may draw an optimistic analogy. At present, we are also learning how human brains engage in scientific cognition, but that does not obviate the need for our scientific communities to continue to generate theories and test them against our unfolding experience. Knowing *how* the brain works so as to generate and constantly improve our scientific understanding will not obviate the need to *keep it* working toward that worthy end, though it may help us to improve our pursuit thereof. Similarly, knowing *how* the brain works to generate and constantly improve our *moral understanding* will not obviate the need to *keep it* working toward that worthy end, though it may help us to improve our pursuit thereof. I will close on this hopeful note.

References

- Campbell, M., Hoane, A. J. & Hsu, F.-H. (2002). Deep blue. *Artificial Intelligence*, 134 (1-2), 57-83.
[10.1016/S0004-3702\(01\)00129-1](https://doi.org/10.1016/S0004-3702(01)00129-1)
- Solomon, R. C. (Ed.) (2001). *Introducing Philosophy*. New York, NY: Oxford University Press.

Applied Metascience of Neuroethics

A Commentary on Paul M. Churchland

Hannes Boelsen

This commentary is the first case study in the applied metascience of neuroethics, that is, the application of a metascientific approach to neuroethical research. I apply a bottom-up approach to neuroethics to Churchland's publication. The bottom-up approach to neuroethics is a quantitative approach (based on scientometric methods) that, among other things, allows us to outline the field from 1995 until 2012 through the development of fifteen subject categories or topic prototypes. Each subject category or topic prototype is defined by up to thirty-one keywords that appear frequently in the abstracts and titles of the publications in the Mainz neuroethics bibliography. The connection strength between two subject categories or topic prototypes depends upon the number of shared publications, that is, the number of publications that can be assimilated to both subject categories or topic prototypes. Accordingly, a keyword-based search of the abstract and title of any publication in neuroethics allows us to assimilate it to (at least) one subject category or topic prototype and, thereby, localize it within neuroethics and reveal its degrees of relevance to neuroethical research, as measured by the connection strengths between the subject categories or topic prototypes. A case study on Churchland's publication led to the following results: the publication is localized in the subject category or topic prototype *Moral Theory*, has high degrees of relevance to research that can be assimilated to the subject categories or topic prototypes *Neuroimaging*, *Philosophy of Mind and Consciousness*, and *Economic and Social Neuroscience*, and has low degrees of relevance to research that can be assimilated to the subject categories or topic prototypes *Addiction*, *Brain Death and Severe Disorders of Consciousness*, *Brain Stimulation*, *Enhancement*, *Legal Studies*, *(Medical) Research and Medicine*, *Molecular Neurobiology and Genetics*, *Neuroscience and Society*, *Neurosurgery*, *Psychiatric and Neurodegenerative Diseases and Disorders*, and *Psychopharmacology*. Such results can be fed back into neuroethical research, which, in turn, can optimize neuroethics itself and, hence, improve our pursuit of moral understanding. The take-home messages are as follows: potential follow-up studies on Churchland's publication should consider my case study results and analysis and, furthermore, future neuroethical research should be more careful to take applied metascience of neuroethics into account. This can be done at different stages of research. If this general idea is on the right track, then applied metascience of neuroethics is complementary to (and perhaps even extends) Churchland's argument, only on a different level.

Keywords

Bibliography | Bibliometrics | Bottom-up | Ethics | Metascience | Neuroethics | Scientometrics | Top-down

1 Introduction

In *Rules: The basis of morality...?*, Churchland points at several problems for classical rule-based accounts of moral knowledge that attempt to identify morally valid behavior-guid-

ing rules and the sources of their authority. Those problems (all based on the fundamental assumption that rules in the literal sense require a language) show that we need a non-

Commentator

[Hannes Boelsen](#)

hboelsen@students.uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Target Author

[Paul M. Churchland](#)

pchurchland@ucsd.edu
University of California
San Diego, CA, U.S.A.

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu
Monash University
Melbourne, Australia

classical non-rule-based account of moral knowledge. Hence, the author proposes an alternative account from computational neuroscience based on “the best hypothesis currently available for how the brain both represents and processes information about the world [...] [and] of how the brain learns” (Churchland this collection, p. 8; emphasis omitted): parallel distributed processing (PDP). In PDP, a neural network embodies a conceptual framework that contains knowledge about the world, that is, a configuration of attractor regions, a family of prototype representations, or, rather, a hierarchy of categories (Churchland 2012, p. 33): against this background, moral knowledge is a configuration of synaptic weights in a neural network. Subsequently, this insight is used to reconceive moral competence, moral conflict, and moral reasoning. Moral competence is the personal level competence to apply sub-personal level knowledge to a moral situation by assimilating it to a prior learned category or prototype. A moral conflict, however, is (at least partly) the consequence of a moral situation that has been assimilated to a category or prototype of which it is not an instance. In short, the fallibility of moral cognition leads to competing interpretations of a moral situation and thereby to a disagreement with others. Accordingly, moral reasoning is (at least mostly) not about rules and the sources of their authority but about adequate assimilation of a moral situation to a category or prototype in the first place. Finally, the author concludes: “[k]nowing how the brain works to generate and constantly improve our moral understanding will not obviate the need to keep it working toward that worthy end, though it may help us to improve our pursuit thereof” (Churchland this collection, p. 13; emphasis omitted).

Churchland’s publication has my full support. I agree with what he says, as I do with his general approach. What follows is a complementary (and perhaps even extending) attempt to improve our pursuit of moral understanding, only on a different level: applied metascience of neuroethics (NE), that is, the application of a

metascientific approach to neuroethical research.¹

In this commentary, I apply the (as-yet unpublished) bottom-up approach to NE² offered by Hildt et al. (forthcoming)³ to Churchland’s publication. Thereby, I attempt to achieve my epistemic goal, which is both to localize the publication within NE and reveal its degrees of relevance⁴ to neuroethical research; as well as my argumentative goal, which is to demonstrate that applied metascience of NE can optimize NE itself and, hence, improve our pursuit of moral understanding.

In the following, I introduce NE and present three typical examples of (disadvantageous) contemporary top-down approaches to NE. I then introduce a bottom-up approach to NE. Following this, I apply the bottom-up approach to NE to Churchland’s publication and present my case study results. After this, I analyze my case study results. Finally, I conclude with some suggestions for future research.

2 Top-down approaches to neuroethics

NE, as a combination of applied ethics⁵ and neurophilosophy⁶ (Hildt 2012, p. 11), is an interdisciplinary field at the intersection of neuroscience, medicine, and philosophy that deals with philosophical, ethical, anthropological, and socio-cultural issues related to neuroscience (Metzinger 2012, p. 36). In 2002, this versatile field emerged in the wake of several US-American conferences that were products of the *Zeitgeist*, that is, the Decade of the Brain from 1990 to 1999 (Hildt

1 In general, applied metascience is not limited to NE, but can be performed with any kind of scientific discipline.

2 A bottom-up approach to NE is data-driven, whereas a top-down approach to NE is definition-seeking.

3 I would like to thank my colleagues in the Mainz Research Group on Neuroethics/Neurophilosophy for providing me the opportunity to use the bottom-up approach to NE for the purpose of this commentary.

4 The degrees of relevance of publications to neuroethical research, as measured by the connection strengths between the subject categories or topic prototypes, indicate the probabilities that publications will prove fruitful for neuroethical research.

5 NE is neither another branch of applied ethics (Levy 2011, p. 3) nor reducible to medicine ethics, bioethics, or a subfield thereof. Nevertheless, there is much overlap (Hildt 2012, pp. 11–12) between these fields.

6 Neurophilosophy is a naturalistic and reductive approach towards a unified theory of the mind-brain that requires detailed knowledge about neuroscience (Walter 2013, p. 133).

2012, p. 9). In particular, it is common to identify the dawn of NE with a conference that was held in San Francisco on May 13th and 14th, 2002: *Neuroethics: Mapping the Field* (Marcus 2002). Before this, “most people saw no need for any such field” (Levy 2007, p. 1), but the aforementioned issues came to be perceived as far more important at this time. Nevertheless, we should ask: what exactly is NE? Alongside the first approximation given above, I present three typical examples of contemporary top-down approaches to NE (which I don’t claim to be exhaustive).

From a knowledge-driven perspective (Racine 2008, p. 33), Roskies divides NE into two divisions: the ethics of neuroscience and the neuroscience of ethics. According to Levy, the former “seeks to develop an ethical framework for regulating the conduct of neuroscientific enquiry and the application of neuroscientific knowledge to human beings [...] [whereas the latter studies] the impact of neuroscientific knowledge upon our understanding of ethics itself” (Levy 2007, p. 1). Furthermore:

the ethics of neuroscience can be roughly subdivided into [...] (1) the ethical issues and considerations that should be raised in the course of designing and executing neuroscientific studies and (2) evaluation of the ethical and social impact that the results of those studies might have, or ought to have, on existing social, ethical, and legal structures. (Roskies 2002, p. 21)

This top-down approach to NE emphasizes the philosophical challenges posed by neuroscience (Racine 2008, p. 34), for example, for “philosophical notions such as free-will, self-control, personal identity, and intention” (Roskies 2002, p. 22).

From a technology-driven perspective (Racine 2008, p. 33), Wolpe identifies NE with “both research and clinical applications of neurotechnology, as well as social and policy issues attendant to their use. [...] [Thus, it is] a content field, defined by the technologies it examines rather than any particular philosophical approach” (Wolpe 2004, p. 1894). This top-down approach to NE emphasizes the ethical challenges of using neurotechnology (Racine 2008, p. 33), for ex-

ample, in healthcare and social practices (Racine 2008, p. 32).

From a healthcare-driven perspective (Racine 2008, p. 33), Racine & Illes (2008) propose a definition of NE that “profiles the field as at the intersection of neuroscience and bioethics defined by a general practical goal, that of improving patient care for specific patient populations” (Racine 2008, p. 34). This top-down approach to NE emphasizes the field as “both a scholarly and practical endeavor, akin to medicine, which attempts to understand and intervene” (Racine 2008, p. 34).

In sum, each of the three top-down approaches to NE comprises (despite their convergences) different issues in different subject categories or topic prototypes with different relations to each other. Seemingly, there are as many top-down approaches to NE as philosophers in the field (e.g., Farah 2012⁷; Gazzaniga 2005;⁸ Giordano n. d.;⁹ Moreno 2003;¹⁰ Safire 2007¹¹) but probably even more.¹²

This unsystematic versatility is disadvantageous for any attempt at a precise localization of Churchland’s publication within the field because it suggests that the aforementioned top-down approaches to NE are necessarily incomplete or even inconsistent. Hence, their application can lead to unsatisfactory results—for example, a localization of the publication that depends more on a research agenda than on facts.¹³ The bottom-up approach to NE attempts to provide a solution to this problem.

⁷ Farah characterizes NE as “a broad range of ethical, legal, and social issues raised by progress in neuroscience” (2012, p. 572).

⁸ Gazzaniga understands NE as “the examination of how we want to deal with the social issues of disease, normality, mortality, lifestyle, and the philosophy of living, informed by our understanding of underlying brain mechanisms” (2005, p. xv).

⁹ Giordano identifies NE with “(1) the study of neurological bases of moral cognition, sense and action[,] (2) the field of study that addresses the moral issues that arise in and from neuroscientific research and the clinical practices and social effects/implications that evolve from these investigations[, and] (3) the reciprocal interaction(s) between neurological research/clinical practices and other ethically relevant areas of biomedical sciences” (Giordano n. d.).

¹⁰ Moreno argues that NE “is in some ways old wine in a new bottle” (2003, p. 153).

¹¹ Safire defines NE as “the examination of what is right and wrong, good and bad about the treatment of, perfection of, and welcome invasion or worrisome manipulation of the human brain” (2007, p. 8).

¹² Buniak et al. “provide an iterative, four-part document that affords a repository of international papers, books, and chapters that address the field in overview, and present discussion(s) of more particular aspects and topics of neuroethics” (2014, p. 3).

3 A bottom-up approach to neuroethics

The bottom-up approach to NE is a quantitative approach (based on scientometric methods) that, among other things, allows us to outline the field from 1995 until 2012 through the development of subject categories or topic prototypes.¹⁴ Although similar work has been done before, for example, by Gooray & Ferguson (2013), Garnet et al. (2011), or Seixas & Basto (2008), no bottom-up approach to NE based on such a comprehensive database as that of Hildt et al. (forthcoming) has yet been attempted.¹⁵ To be more precise, they use the Mainz NE bibliography.¹⁶

The Mainz bibliography (launched in 2006) is an open-access online bibliography compiled and provided by the Mainz Research Group on Neuroethics/Neurophilosophy.¹⁷ Currently, the bibliography, as a multimodal compilation of NE publications (e.g., anthologies, edited volumes, journal articles, and monographs), contains about 4095 entries produced between 1949 and mid-2014. On the one hand, the bibliography is based on regular scans of relevant journals from neuroscience and medicine (e.g., *Cortex*, *Der Nervenarzt*, *EMBO Reports*, *Journal of Neurology*, *Journal of the American Medical Association*, *Nature*, *Nature Neuroscience*, *Nature Reviews Neuroscience*, *Neurocritical Care*, *NeuroImage*, *Neurology*, *Neuropsychology Review*, *Psychopharmacology*, *Science*, and *Trends in Cognitive Sciences*),

philosophy (e.g., *American Journal of Bioethics Neuroscience*, *American Journal of Bioethics Bioethics*, *Cambridge Quarterly of Healthcare Ethics*, *Consciousness and Cognition*, *Journal of Applied Philosophy*, *Journal of Medical Ethics*, *Medicine, Health Care and Philosophy*, *Neuroethics*, *Philosophy*, *Psychiatry*, & *Psychology*, *Science and Engineering Ethics*, *Hastings Center Report*, *The Journal of Law, Medicine & Ethics*, and *Theoretical Medicine and Bioethics*), the humanities, and social sciences.¹⁸ On the other hand, the bibliography is based on regular searches of both relevant citation (meta-)databases such as *Web of Science*,¹⁹ *PubMed*,²⁰ and *Scopus*,²¹ and relevant bibliographies such as the *Brainstorm*²² newsletter of the Canadian Neuroethics and Mental Health Interest Group. The bibliography also incorporates irregular additions of relevant publications (mainly anthologies, edited volumes, and monographs) as soon as the Research Group on Neuroethics/Neurophilosophy becomes aware of them. Regarding the selection criteria for publications from the various sources, publications from neuroscience or medicine are selected if they refer to the philosophical, ethical, anthropological, or socio-cultural impact of the presented results,

¹³ For example, facts that are necessary for an adequate mapping of the field may have been (un-)intentionally overlooked.

¹⁴ In Hildt et al. (forthcoming), the developed subject categories or topic prototypes form the basis for further scientometric analysis of the data. For example, the subject categories or topic prototypes allow us to examine the development and institutionalization of NE (e.g., temporal development, structure and disciplinary institutionalization, and reciprocal shaping of NE and related disciplines).

¹⁵ For example, Gooray & Ferguson's (2013) database contains about 205 entries dating from 2000 to 2012. The database is based on books and articles from the following twelve journals: *Neuroethics*, *American Journal of Bioethics Neuroscience*, *Nature Reviews Neuroscience*, *Annual Review of Neuroscience*, *Behavioral and Brain Sciences*, *Molecular Psychiatry*, *Nature Neuroscience*, *Neuron*, *Trends in Neurosciences*, *Frontiers in Neuroendocrinology*, *Annals of Neurology*, and *Progress in Neurobiology*. In contrast, Hildt et al.'s (forthcoming) database contains about 2296 entries dating from 1995 to 2012. It is based on books and articles from more than 700 journals.

¹⁶ <https://teamweb.uni-mainz.de/fb05/Neuroethics/SitePages/Home.aspx>

¹⁷ <http://www.blogs.uni-mainz.de/fb05philosophieengl/further-institutions/research-group-on-neuroethics-and-neurophilosophy/>

¹⁸ The aforementioned selection of twenty-nine journals comprises those journals that had added at least twenty publications to the Mainz NE bibliography before mid-2014. The number of publications ranges from (at least) 352 publications (*American Journal of Bioethics Neuroscience*), 298 publications (*American Journal of Bioethics*), 211 publications (*Neuroethics*), 91 publications (*Nature Reviews Neuroscience*), 68 publications (*Der Nervenarzt*), 61 publications (*Nature Neuroscience*), 58 publications (*Journal of Medical Ethics*), 57 publications (*Journal of Neurology*), 54 publications (*Nature and Neurology*), 46 publications (*Bioethics*), 40 publications (*NeuroImage and Science and Engineering Ethics*), 37 publications (*Trends in Cognitive Sciences*), 35 publications (*Hastings Center Report*), 31 publications (*Journal of the American Medical Association*, *Medicine, Health Care and Philosophy*, and *Philosophy, Psychiatry, & Psychology*), 28 publications (*Science*), 26 publications (*Cortex and EMBO Reports*), 23 publications (*Neurocritical Care and Neuropsychology Review*), 22 publications (*Cambridge Quarterly of Healthcare Ethics*, *Journal of Applied Philosophy*, and *Psychopharmacology*), 21 publications (*The Journal of Law, Medicine & Ethics*) to 20 publications (*Consciousness and Cognition* and *Theoretical Medicine and Bioethics*). This selection of twenty-nine journals could be a fruitful starting point for future scientometric research related to NE. Besides this, the Mainz NE bibliography comprises journals that have added less than twenty publications (e.g., *Behavioral and Brain Sciences* and *Philosophical Psychology*).

¹⁹ <http://www.webofknowledge.com>

²⁰ <http://www.ncbi.nlm.nih.gov/pubmed>

²¹ <http://www.scopus.com>

²² <http://www.ircm.qc.ca/LARECHERCHE/AXES/NEURO/NEURO-ETHIQUE/PAGES/GROUPE.ASPX?PFLG=1033&lan=1033>

whereas publications from philosophy, the humanities, or social sciences are selected if they refer to empirical results from neuroscience or medicine. Moreover, non-transdisciplinary publications are selected if the Research Group on Neuroethics/Neurophilosophy considers them to be relevant to NE.²³

Subsequently, Hildt et al. (forthcoming) use a bibliometric analysis of the Mainz NE bibliography from 1995 until 2012 to develop, among other things, fifteen subject categories or topic prototypes on content-based criteria. Thereby, each subject category or topic prototype is defined by up to thirty-one keywords that appear frequently in the abstracts and titles of the publications. These fifteen subject categories or topic prototypes are *Addiction*, *Brain Death and Severe Disorders of Consciousness*, *Brain Stimulation*, *Enhancement*, *Legal Studies*, *(Medical) Research and Medicine*, *Molecular Neurobiology and Genetics*, *Moral Theory*, *Neuroimaging*, *Neuroscience and Society*, *Neurosurgery*, *Philosophy of Mind and Consciousness*, *Psychiatric and Neurodegenerative Diseases and Disorders*, *Psychopharmacology*, and *Social and Economic Neuroscience*. Each subject category or topic prototype represents certain issues discussed in NE²⁴ and, taken together, they outline the field.²⁵ Importantly, Hildt et al. (forthcoming) also determine, among other things, the connection strengths²⁶ between the subject categories or topic prototypes within NE. Due to the content-based development of the subject categories or topic prototypes, a keyword-based search of the abstract and title of any publication in NE allows us to

assimilate it to (at least) one subject category or topic prototype²⁷ and, thereby, localize it within NE.

In the following, I apply the bottom-up approach to NE to Churchland's publication and present my case study results. I thereby attempt to achieve the first part of my epistemic goal, which is to localize Churchland's publication within NE.

4 Case study results

The keyword-based search of the abstract and title of Churchland's publication reveals that it can be assimilated to the subject category or topic prototype *Moral Theory*, that is, a subject category or topic prototype that comprises publications on the psychology and neurobiology of moral-decision making, publications on determinism, free-will, and the function of moral theory in the neurosciences, and publications on challenges to established interpretations of morally significant concepts such as autonomy, responsibility, and human nature.

This subject category or topic prototype has strong connections to the subject categories or topic prototypes *Neuroimaging*, *Philosophy of Mind and Consciousness*, and *Social and Economic Neuroscience*, and weak connections to the subject categories or topic prototypes *Addiction*, *Brain Death and Severe Disorders of Consciousness*, *Brain Stimulation*, *Enhancement*, *Legal Studies*, *(Medical) Research and Medicine*, *Molecular Neurobiology and Genetics*, *Neuroscience and Society*, *Neurosurgery*, *Psychiatric and Neurodegenerative Diseases and Disorders*, and *Psychopharmacology*. The strong connections can be explained by a high number of shared publications, that is, a high number of publications that can be assimilated to both the subject category or topic prototype *Moral Theory* and the subject category or topic prototype *Neuroimaging*, *Philosophy of Mind and Consciousness*, or *Social and Economic Neuros-*

²³ For example, a publication in medicine on the effects of neuroleptics, antidepressants, stimulants, or tranquilizers is selected if it could offer a contribution to the interdisciplinary debate on psychopharmacological cognitive enhancement.

²⁴ Combinations of the subject categories or topic prototypes are able to represent almost every issue discussed in NE.

²⁵ Bottom-up approaches to NE attempt to provide maximally parsimonious bottom-up descriptions of their target phenomenon (e.g., NE as a dynamical publication state-space). If the top-down descriptions of NE, provided by the top-down approaches to NE, are neither identical with nor reducible to the bottom-up descriptions of NE, then, using a superficial analogy to Churchland's eliminative materialism (1981), an interesting question is whether or not (and, if so, which of) the top-down descriptions of NE can be eliminated.

²⁶ The connection strength between two subject categories or topic prototypes depends upon the number of shared publications, that is, the number of publications that can be assimilated to both subject categories or topic prototypes.

²⁷ A publication can be assimilated to a subject category or topic prototype if its abstract and title contain (at least) one of the keywords that define the subject category or topic prototype. As such, less than ten percent of the total publications could not be assimilated to a subject category or topic prototype. An interesting question is whether or not (and, if so, how) those publications can still be regarded as belonging to NE.

cience. The weak connections can be explained by a low number of shared publications, that is, a low number of publications that can be assimilated to both the subject category or topic prototype *Moral Theory* and the subject category or topic prototype *Addiction, Brain Death and Severe Disorders of Consciousness, Brain Stimulation, Enhancement, Legal Studies, (Medical) Research and Medicine, Molecular Neurobiology and Genetics, Neuroscience and Society, Neurosurgery, Psychiatric and Neurodegenerative Diseases and Disorders, or Psychopharmacology* (Hildt et al. forthcoming).

In the following, I analyze my results. I thereby attempt to achieve the second part of my epistemic goal, which is to reveal the degrees of relevance of Churchland's publication to neuroethical research; as well as my argumentative goal, which is to demonstrate that applied metascience of NE can optimize NE itself and, hence, improve our pursuit of moral understanding.

5 Analysis

First of all, the degrees of relevance of publications to neuroethical research are measured by the connection strengths between the subject categories or topic prototypes. The connection strengths between subject categories or topic prototypes depend upon the numbers of shared publications. The numbers of shared publications can be explained by the degrees of overlap of content, methodology, or both. The degrees of overlap of content, methodology, or both, in turn, indicate the probabilities that publications will prove fruitful for neuroethical research. In short, the degrees of relevance of publications to neuroethical research, as measured by the connection strengths between subject categories or topic prototypes, indicate the probabilities that publications will prove fruitful for neuroethical research.

Based on my results, Churchland's publication has high degrees of relevance to research that can be assimilated to the subject categories or topic prototypes *Moral Theory, Neuroimaging, Philosophy of Mind and Consciousness*, or *Social and Economic Neuroscience* because of

the strong connections between the subject category or topic prototype *Moral Theory* and the subject categories or topic prototypes *Neuroimaging, Philosophy of Mind and Consciousness*, and *Social and Economic Neuroscience*. The strong connections can be explained by the high numbers of shared publications. The high numbers of shared publications can be explained by the high degrees of overlap of either content, methodology, or both.²⁸ This, in turn, indicates high probabilities that Churchland's publication will prove fruitful for research that can be assimilated to the aforementioned subject categories or topic prototypes. Conversely, Churchland's publication has low degrees of relevance to research that can be assimilated to the subject categories or topic prototypes *Addiction, Brain Death and Severe Disorders of Consciousness, Brain Stimulation, Enhancement, Legal Studies, (Medical) Research and Medicine, Molecular Neurobiology and Genetics, Neuroscience and Society, Neurosurgery, Psychiatric and Neurodegenerative Diseases and Disorders*, or *Psychopharmacology* because of the weak connections between the subject category or topic prototype *Moral Theory* and the aforementioned subject categories or topic prototypes. Here are some brief theoretical considerations.

Churchland's publication is highly relevant to research that can be assimilated to the subject category or topic prototype *Economic and Social Neuroscience*, suggesting that his idea of reconceiving moral decision-making in terms of PDP could prove fruitful for neuroethical research that refers to the underlying physiology of economic or social decision-making. This application might show that moral, economic, and social decision-making share important properties but differ in others. This possible result could then be fed back into neuroethical research.

Churchland's publication is also highly relevant to research that can be assimilated to the subject category or topic prototype *Neuroima-*

²⁸ Accordingly, publications that can be assimilated to the subject category or topic prototype *Moral Theory, Neuroimaging, Philosophy of Mind and Consciousness*, or *Social and Economic Neuroscience* are highly relevant to the subject of Churchland's publication.

ging, suggesting that his idea of reconceiving moral decision-making in terms of PDP could prove fruitful for neuroethical research that refers to imaging techniques that visualize the brain, such as cranial computed tomography (CCT), electroencephalography (EEG), magnetic resonance imaging (MRI), functional magnetic resonance imaging (fMRI), and positron emission tomography (PET) (Hildt 2012, p. 11). For example, it could be used to reconceive the classic distinction between off-track and truth-tracking processes in genealogical debunking arguments²⁹ that refer to fMRI research (e.g., Greene 2008 and Singer 2005). This application might show that the classic distinction is neurobiologically implausible, which would mean that arguments relying on this distinction are implausible as well. This possible result could then be fed back into neuroethical research.

Moreover, the possible (yet unrecognized) relevance of Churchland's publication to research that can be assimilated to the subject categories or topic prototypes *Addiction*, *Brain Death and Severe Disorders of Consciousness*, *Brain Stimulation*, *Enhancement*, *Legal Studies*, *(Medical) Research and Medicine*, *Molecular Neurobiology and Genetics*, *Neuroscience and Society*, *Neurosurgery*, *Psychiatric and Neurodegenerative Diseases and Disorders*, and *Psychopharmacology* could have been emphasized more strongly by including keywords in the abstract and title that define the aforementioned subject categories or topic prototypes, which, in turn, could have increased the connection strengths between those subject categories or topic prototypes and the subject category or topic prototype *Moral Theory*. A possible outcome of this could have been the revelation of a systematic overlap of content, methodology, or both that has been neglected so far. And this possible result could then have been fed back into neuroethical research.³⁰

This feedback process, in turn, can optimize NE itself and, hence, improve our pursuit of moral understanding because it can help to “produce better ethical theories [...] and contribute toward the great project of better understanding ourselves” (Levy 2011, p. 8). Apparently, a recurring pattern emerges: the bottom-up approach to NE can be applied to neuroethical research, which, in turn, can lead to such results that can be fed back into it, which, in turn, can optimize NE itself and, hence, improve our pursuit of moral understanding.

6 Concluding remarks

In this commentary, I applied the bottom-up approach to NE to Churchland's publication. I thereby attempted to localize the publication within NE and reveal its degrees of relevance to neuroethical research, and to demonstrate that applied metascience of NE can optimize NE itself and, hence, improve our pursuit of moral understanding.

Assuming that I have achieved the former, which was my epistemic goal, the first and more specific take-home message is that potential follow-up studies on Churchland's publication should consider my case study results and analysis, that is, they should both bring together research that can be assimilated to the subject categories or topic prototypes *Moral Theory*, *Neuroimaging*, *Philosophy of Mind and Consciousness*, and *Social and Economic Neuroscience*, and build bridges to research that can be assimilated to the subject categories or topic prototypes *Addiction*, *Brain Death and Severe Disorders of Consciousness*, *Brain Stimulation*, *Enhancement*, *Legal Studies*, *(Medical) Research and Medicine*, *Molecular Neurobiology and Genetics*, *Neuroscience and Society*, *Neurosurgery*, *Psychiatric and Neurodegenerative Diseases and Disorders*, and *Psychopharmacology*. Assuming that I have achieved the latter, which was my argumentative goal, the more general take-home message is that future neuroethical research should be more careful to take applied metascience of NE into account because it can optimize NE itself and, hence, improve our pursuit of moral understanding.

²⁹ According to Kahane, the general form of a genealogical debunking argument is the following: S's belief that p is explained by x. But, x is an off-track process, that is, not a truth-tracking process, with respect to p. Therefore, S's belief that p is unjustified (Kahane 2011, p. 106).

³⁰ Of course, this theoretical consideration is not meant to be a serious criticism of Churchland's publication, because the bottom-up approach to NE was not available to him at the time of writing. It rather shows that applied metascience of NE can help us discover new pathways and directions for future neuroethical research.

In the case of the bottom-up approach to NE, this can be done at different stages of research. First, while seeking inspiration for research, researchers and students can bypass well-trodden paths in NE and identify (as yet) unorthodox ones from the very beginning. Second, while pursuing these (or already well-trodden) paths, scholars can optimize the efficiency of their own research. Third, while preparing their research for publication, they can prepare abstracts and titles in such a manner as to optimally reflect the publications' (real or intended) degrees of relevance to specific subject categories or topic prototypes. Fourth and finally, when taking it into account, they shape NE in such a way that it provides input for more fine-grained follow-up models in the metascience of NE.

If this general idea is on the right track, then applied metascience of NE is complementary to (and perhaps even extends) Churchland's argument, only on a different level: "knowing how the brain works to generate and constantly improve our moral understanding will not obviate the need to keep it working towards that worthy end" (Churchland this collection, p. 13; emphasis omitted), just as knowing how to optimize NE will not do this either, though both "may help us to improve our pursuit thereof" (Churchland this collection, p. 13). Only time will tell.

References

- Buniak, L., Darragh, M. & Giordano, J. (2014). A four-part working bibliography of neuroethics: part 1: Overview and reviews – defining and describing the field and its practices. *Philosophy, Ethics, and Humanities in Medicine*, 9 (9), 1-14. [10.1186/1747-5341-9-9](https://doi.org/10.1186/1747-5341-9-9)
- Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy*, 78 (2), 67-90. [10.2307/2025900](https://doi.org/10.2307/2025900)
- (2012). *Plato's camera: How the physical brain captures a landscape of abstract universals*. Cambridge, MA: MIT Press.
- (2015). Rules: The basis of morality...? In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Farah, M. (2012). Neuroethics: The ethical, legal, and societal impact of neuroscience. *Annual Review of Psychology*, 63 (1), 571-591. [10.1146/annurev.psych.093008.100438](https://doi.org/10.1146/annurev.psych.093008.100438)
- Garnet, A., Whiteley, L., Piwowar, H., Rasmussen, E. & Illes, J. (2011). Neuroethics and fMRI: Mapping a fledgling relationship. *PLoS ONE*, 6 (4), e18537. [10.1371/journal.pone.0018537](https://doi.org/10.1371/journal.pone.0018537)
- Gazzaniga, M. (2005). *The ethical brain*. New York, NY: Dana Press.
- Giordano, J. (2014). Neuroethics. *At the intersection of neuroscience, morality, and society*, Retrieved September 11, 2014, from <http://www.neurobioethics.org>
- Gooray, E. & Ferguson, C. (2013). Neuroethics as a field: How much has it grown, about what, and by whom? *Unpublished manuscript*, Retrieved July 15, 2014, from <http://www.neuroethicssociety.org/survey-neuroethics-as-a-field>
- Greene, J. (2008). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.) *Moral psychology: The neuroscience of morality: Emotion, brain disorders, and development* (pp. 35-79). Cambridge, MA: MIT Press.
- Hildt, E. (2012). *Neuroethik*. München, GER: Reinhardt.
- Hildt, E., Leefmann, J. & Levallois, C. (forthcoming). A bottom-up analysis of "neuroethics".
- Kahane, G. (2011). Evolutionary debunking arguments. *Noûs*, 45 (1), 103-125. [10.1111/j.1468-0068.2010.00770.x](https://doi.org/10.1111/j.1468-0068.2010.00770.x)
- Levy, N. (2007). *Neuroethics: Challenges for the 21st century*. Cambridge, UK: Cambridge University Press.
- (2011). Neuroethics: A new way of doing ethics. *American Journal of Bioethics Neuroscience*, 2 (2), 3-9. [10.1080/21507740.2011.557683](https://doi.org/10.1080/21507740.2011.557683)

- Marcus, S. (Ed.) (2002). *Neuroethics: Mapping the field*. New York, NY: Dana Press.
- Metzinger, T. (2012). Zehn Jahre Neuroethik des pharmazeutischen kognitiven Enhancements: Aktuelle Probleme und Handlungsrichtlinien für die Praxis. *Fortschritte der Neurologie und Psychiatrie*, 80 (1), 36-43. [10.1055/s-0031-1282051](https://doi.org/10.1055/s-0031-1282051)
- Moreno, J. (2003). Neuroethics: An agenda for neuroscience and society. *Nature Reviews Neuroscience*, 4 (2), 149-153. [10.1038/nrn1031](https://doi.org/10.1038/nrn1031)
- Racine, E. (2008). *Pragmatic neuroethics: Improving treatment and understanding of the mind-brain*. Cambridge, MA: MIT Press.
- Racine, E. & Illes, J. (2008). Neuroethics. In P. Singer & A. Viens (Eds.) *Cambridge textbook of bioethics* (pp. 495-503). Cambridge, UK: Cambridge University Press.
- Roskies, A. (2002). Neuroethics for the new millenium. *Neuron*, 35 (1), 21-23. [10.1016/S0896-6273\(02\)00763-8](https://doi.org/10.1016/S0896-6273(02)00763-8)
- Safire, W. (2007). Visions for a new field of “neuroethics”. In W. Glannon (Ed.) *Defining right and wrong in brain science: essential readings in neuroethics* (pp. 7-11). New York, NY: Dana Press.
- Seixas, D. & Basto, M. (2008). Ethics in fMRI studies. A review of the EMBASE and MEDLINE literature. *Clinical Neuroradiology*, 18 (2), 79-87. [10.1007/s00062-008-8009-5](https://doi.org/10.1007/s00062-008-8009-5)
- Singer, P. (2005). Ethics and intuitions. *The Journal of Ethics*, 9 (3-4), 331-352. [10.1007/s10892-005-3508-y](https://doi.org/10.1007/s10892-005-3508-y)
- Walter, H. (2013). Neurophilosophie und Philosophie der Neurowissenschaft. In A. Stephan & S. Walter (Eds.) *Handbuch Kognitionswissenschaft* (pp. 133-138). Stuttgart, GER: Metzler.
- Wolpe, P. (2004). Neuroethics. *Encyclopedia of bioethics*. 3rd ed. Vol. 4 (pp. 1894-1898). New York, NY: Macmillan Reference.

A Skeptical Note on Bibliometrics

A Reply to Hannes Boelsen

Paul M. Churchland

Author

Paul M. Churchland
pchurchland@ucsd.edu
University of California
San Diego, CA, U.S.A.

Commentator

Hannes Boelsen
hboelsen@students.uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Editors

Thomas Metzinger
metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

My thanks to Boelsen for his penetrating understanding of my modest contribution to this collection, and for placing its significance in a much broader context, namely, the context of the full range of scientific and philosophical research to which it might be *relevant*. Indeed, his principal topic is the emerging internet mechanism for evaluating the relevance of *any* publication to the research interests of scholars in general, a mechanism that allows a specific scholar to identify, from among the teeming multitude, exactly those published papers most likely to be of interest to him or her. Its brief application to my own paper in this collection is just one illustration of its wide-ranging *possible* applications.

The mechanism he describes – namely, the calculation of “connections strengths” between the prototype topics and the key words found in the abstracts of any arbitrarily chosen pair of publications – is an interesting elaboration of the simpler “key words” convention already in widespread use in modern journals, a convention that has already proven to be very useful to scholars all across the academic spectrum, as we all know. Taking the variable “connection strengths” – as defined by Boelsen – between those already-salient indexes into account, and making them systematically available also, would seem only to enhance the usefulness of the mechanisms already in play.

And no doubt it would. However, and its undoubted advantages conceded, there is an unfortunate limit on the usefulness of such a mechanism, a limit already familiar to us from our experience with the existing conventions of abstracts and key words. They are intellectually useful only if, and only to the extent that, one *already understands* the “key words” involved, and the research areas that they name. Otherwise, the mechanism here at issue does no more than *cluster together* distinct publications as having “the same”, or “closely similar”, intellectual concerns. That is, it does provide a map of the “topical concentrations” at the presumptive current “ceiling” of academic understanding, but it does not itself raise the “level” of that ceiling. By itself, it provides no novel or additional understanding of the various topics themselves displayed in its many lists. *That* sort of achievement, if it is realized at all, must be made by those occasional thinkers who actually *read* the papers thus clustered together, and subsequently manage to *solve* one or more of the problems that they still leave open, by using the quite different mechanisms that reside within the human *brain*.

In sum, the mechanism described by Boelsen will certainly help aspiring scholars to *catch up* on the already existing research that is relevant to their own research interests, and may thereby stimulate further research. But any intellectual or theoretical novelties will have to come from the subsequent researches of those aspiring scholars themselves, and not from the mechanism described by Boelsen. That said, in constructing “key-word lists” for my own papers in the future, I will keep the mechanism described by Boelsen firmly in mind. And for a reason that would not have occurred to me, save for Boelsen’s commentary. In constructing the abstract and key-words list for my own paper in this collection, I did not pay special attention to the possible *novel uses* to which its contents might be put, and the possible *novel topics* for which it might provide enlightenment. To illustrate this point, I would now include the key words *moral pathology*, *moral character*, *moral reasoning*, *moral development*, and *moral conflict* in such a list. For this belated oppor-

tunity, here on this page, I am once again in Boelsen’s debt.

Embodied Prediction

Andy Clark

Versions of the “predictive brain” hypothesis rank among the most promising and the most conceptually challenging visions ever to emerge from computational and cognitive neuroscience. In this paper, I briefly introduce (section 1) the most radical and comprehensive of these visions—the account of “active inference”, or “action-oriented predictive processing” (Clark 2013a), developed by Karl Friston and colleagues. In section 2, I isolate and discuss four of the framework’s most provocative claims: (i) that the core flow of information is top-down, not bottom-up, with the forward flow of sensory information replaced by the forward flow of prediction error; (ii) that motor control is just more top-down sensory prediction; (iii) that efference copies, and distinct “controllers”, can be replaced by top-down predictions; and (iv) that cost functions can fruitfully be replaced by predictions. Working together, these four claims offer a tantalizing glimpse of a new, integrated framework for understanding perception, action, embodiment, and the nature of human experience. I end (section 3) by sketching what may be the most important aspect of the emerging view: its ability to embed the use of fast and frugal solutions (as highlighted by much work in robotics and embodied cognition) within an over-arching scheme that includes more structured, knowledge-intensive strategies, combining these fluently and continuously as task and context dictate.

Keywords

Active inference | Embodied cognition | Motor control | Prediction | Prediction error

Author

Andy Clark

andy.clark@ed.ac.uk

University of Edinburgh
Edinburgh, United Kingdom

Commentator

Michael Madary

madary@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Mind turned upside down?

PP (Predictive processing; for this terminology, see Clark 2013a) turns a traditional picture of perception on its head. According to that once-standard picture (Marr 1982), perceptual processing is dominated by the forward flow of information transduced from various sensory receptors. As information flows forward, a progressively richer picture of the real-world scene is constructed. The process of construction would involve the use of stored knowledge of various kinds, and the forward flow of information was subject to modulation and nuancing by top-down (mostly attentional) effects. But the basic picture remained one in which perception was fundamentally a process of “bottom-up feature detection”. In Marr’s theory of vision, detected intensities (arising from surface discon-

tinuities and other factors) gave way to detected features such as blobs, edges, bars, “zero-crossings”, and lines, which in turn gave way to detected surface orientations leading ultimately (though this step was always going to be problematic) to a three-dimensional model of the visual scene. Early perception is here seen as building towards a complex world model by a feedforward process of evidence accumulation. Traditional perceptual neuroscience followed suit, with visual cortex (the most-studied example) being “traditionally viewed as a hierarchy of neural feature detectors, with neural population responses being driven by bottom-up stimulus features” (Egner et al. 2010, p. 16601). This was a view of the perceiving brain as passive and stimulus-driven, taking energetic inputs

from the senses and turning them into a coherent percept by a kind of step-wise build-up moving from the simplest features to the more complex: from simple intensities up to lines and edges and on to complex meaningful shapes, accumulating structure and complexity along the way in a kind of Lego-block fashion.

Such views may be contrasted with the increasingly active views that have been pursued over the past several decades of neuroscientific and computational research. These views (Ballard 1991; Churchland et al. 1994; Ballard et al. 1997) stress the active search for task-relevant information just-in-time for use. In addition, huge industries of work on intrinsic neural activity, the “resting state” and the “default mode” (for a review, see Raichle & Snyder 2007) have drawn our attention to the ceaseless buzz of neural activity that takes place even in the absence of ongoing task-specific stimulation, suggesting that much of the brain’s work and activity is in some way ongoing and endogenously generated.

Predictive processing plausibly represents the last and most radical step in this retreat from the passive, input-dominated view of the flow of neural processing. According to this emerging class of models, naturally intelligent systems (humans and other animals) do not passively await sensory stimulation. Instead, they are constantly active, trying to predict the streams of sensory stimulation before they arrive. Before an “input” arrives on the scene, these pro-active cognitive systems are already busy predicting its most probable shape and implications. Systems like this are already (and almost constantly) poised to act, and all they need to process are any sensed deviations from the predicted state. It is these calculated deviations from predicted states (known as *prediction errors*) that thus bear much of the information-processing burden, informing us of what is salient and newsworthy within the dense sensory barrage. The extensive use of top-down probabilistic prediction here provides an effective means of avoiding the kinds of “representational bottleneck” feared by early opponents (e.g., Brooks 1991) of representation-heavy—but feed-forward dominated—forms of pro-

cessing. Instead, the downward flow of prediction now does most of the computational “heavy-lifting”, allowing moment-by-moment processing to focus only on the newsworthy departures signified by salient (that is, high-precision—see section 3) prediction errors. Such economy and preparedness is biologically attractive, and neatly sidesteps the many processing bottlenecks associated with more passive models of the flow of information.

Action itself (more on this shortly) then needs to be reconceived. Action is not so much a response to an input as a neat and efficient way of selecting the next “input”, and thereby driving a rolling cycle. These hyperactive systems are constantly predicting their own upcoming states, and actively moving so as to bring some of them into being. We thus act so as to bring forth the evolving streams of sensory information that keep us viable (keeping us fed, warm, and watered) and that serve our increasingly recondite ends. PP thus implements a comprehensive reversal of the traditional (bottom-up, forward-flowing) schema. The largest contributor to ongoing neural response, if PP is correct, is the ceaseless anticipatory buzz of downwards-flowing neural prediction that drives both perception and action. Incoming sensory information is just one further factor perturbing those restless pro-active seas. Within those seas, percepts and actions emerge via a recurrent cascade of sub-personal predictions forged (see below) from unconscious expectations spanning multiple spatial and temporal scales.

Conceptually, this implies a striking reversal, in that the driving sensory signal is really just providing corrective feedback on the emerging top-down predictions.¹ As ever-active prediction engines, these kinds of minds are not, fundamentally, in the business of solving puzzles given to them as inputs. Rather, they are in the business of keeping us one step ahead of the game, poised to act and actively eliciting the sensory flows that keep us viable and fulfilled. If this is on track, then just about every aspect of the passive forward-flowing model is false. We are not passive cognitive couch potatoes so

¹ For this observation, see Friston (2005), p. 825, and the discussion in Hohwy (2013).

much as proactive predictors, forever trying to stay one step ahead of the incoming waves of sensory stimulation.

2 Radical predictive processing

Such models involve a number of quite radical claims. In the present treatment, I propose focusing upon just four:

1. The core flow of information is top-down, not bottom-up, and the forward flow of sensory information is replaced by the forward flow of prediction error.
2. Motor control is just more top-down sensory prediction.
3. Efference copies, and distinct “controllers” (inverse models) are replaced by top-down predictions.
4. Cost functions are absorbed into predictions.

One thing I shan’t try to do here is rehearse the empirical evidence for the framework. That evidence (which is substantial but importantly incomplete) is rehearsed in [Clark \(2013a\)](#) and [Hohwy \(2013, this collection\)](#). For a recent attempt to specify a neural implementation, see [Bastos et al. \(2012\)](#). I now look at each of these points in turn:

2.1 The core flow of information is top-down, not bottom-up, and the forward flow of sensory information is replaced by the forward flow of prediction error

This is the heart and soul of the radical vision. Incoming sensory information, if PP is correct, is constantly met with a cascade of top-down prediction, whose job is to predict the incoming signal across multiple temporal and spatial scales.

To see how this works in practice, consider a seminal proof-of-concept by [Rao & Ballard \(1999\)](#). In this work, prediction-based learning targets image patches drawn from natural scenes using a multi-layer artificial neural network. The network had no pre-set task apart from that of using the downwards connections

to match input samples with successful predictions. Instead, visual signals were processed via a hierarchical system in which each level tried (in the way just sketched) to predict activity at the level below it using recurrent (feedback) connections. If the feedback successfully predicted the lower-level activity, no further action was required. Failures to predict enabled tuning and revision of the model (initially, just a random set of connection weights) generating the predictions, thus slowly delivering knowledge of the regularities governing the domain. In this architecture, forward connections between levels carried only the “residual errors” ([Rao & Ballard 1999](#), p. 79) between top-down predictions and actual lower level activity, while backward or recurrent connections carried the predictions themselves.

After training, the network developed a nested structure of units with simple-cell-like receptive fields and captured a variety of important, empirically-observed effects. One such effect was “end-stopping”. This is a “non-classical receptive field” effect in which a neuron responds strongly to a short line falling within its classical receptive field but (surprisingly) shows diminishing response as the line gets longer. Such effects (and with them, a whole panoply of “context effects”) emerge naturally from the use of hierarchical predictive processing. The response tails off as the line gets longer, because longer lines and edges were the statistical norm in the natural scenes to which the network was exposed in training. After training, longer lines are thus what is first predicted (and fed back, as a hypothesis) by the level-two network. The strong firing of some level-one “edge cells”, when they are driven by shorter lines, thus reflects not successful feature detection by those cells but rather error or mismatch, since the short segment was not initially predicted by the higher-level network. This example neatly illustrates the dangers of thinking in terms of a simple cumulative flow of feature-detection, and also shows the advantages of re-thinking the flow of processing as a mixture of top-down prediction and bottom-up error correction.² In ad-

² This does not mean that there are no cells in v1 or elsewhere whose responses match the classical profile. PP claims that each neural area

dition it highlights the way these learning routines latch on to the world in a manner specified by the training data. End-stopped cells are simply a response to the structure of the natural scenes used in training, and reflect the typical length of the lines and edges in these natural scenes. In a very different world (such as the underwater world of some sea-creatures) such cells would learn very different responses.

These were early and relatively low-level results, but the predictive processing model itself has proven rich and (as we shall see) widely applicable. It assumes only that the environment generates sensory signals by means of nested interacting causes and that the task of the perceptual system is to invert this structure by learning and applying a structured internal model—so as to predict the unfolding sensory stream. Routines of this kind have recently been successfully applied in many domains, including speech perception, reading, and recognizing the actions of oneself and of other agents (see [Poepel & Monahan 2011](#); [Price & Devlin 2011](#); [Friston et al. 2011](#)). This is not surprising, since the underlying rationale is quite general. If you want to predict the way some set of sensory signals will change and evolve over time, a good thing to do is to learn how those sensory signals are determined by interacting external causes. And a good way to learn about those interacting causes is to try to predict how the sensory signal will change and evolve over time.

Now try to imagine this this on a very grand scale. To predict the visually presented scene, the system must learn about edges, blobs, line segments, shapes, forms, and (ultimately) objects. To predict text, it must learn about interacting “hidden” causes in the linguistic domain: causes such as sentences, words, and letters. To predict all of our rich multimodal plays of sensory data, across many scales of space and time, it must learn about interacting hidden causes such as tables, chairs, cats, faces, people, hurricanes, football games, goals,

contains two kinds of cell, or at least supports two functionally distinct response profiles, such that one functionality encodes error and the other current best-guess content. This means that there can indeed be (as single cell recordings amply demonstrate) recognition cells in each area, along with the classical response profiles. For more on this important topic, see [Clark \(2013a\)](#).

meanings, and intentions. The structured world of human experience, if this is correct, comes into view only when all manner of top-down predictions meet (and “explain away”) the incoming waves of sensory information. What propagates forwards (through the brain, away from the sensory peripheries) is then only the mismatches, at every level, with predicted activity.

This makes functional sense. Given that the brain is ever-active, busily predicting its own states at many levels, all that matters (that is, all that is newsworthy, and thus ought to drive further processing) are the incoming surprises: unexpected deviations from what is predicted. Such deviations result in prediction errors reflecting residual differences, at every level and stage of processing, between the actual current signal and the predicted one. These error signals are used to refine the prediction until the sensory signal is best accommodated.

Prediction error thus “carries the news”, and is pretty much the hero (or anti-hero) of this whole family of models. So much so, that it is sometimes said that:

In predictive coding schemes, sensory data are replaced by prediction error, because that is the only sensory information that has yet to be explained. ([Feldman & Friston 2010](#), p. 2)

We can now savor the radicalism. Where traditional, feed-forward-based views see a progressive (though top-down modulated) flow of complex feature-detection, the new view depicts a progressive, complex flow of feature prediction. The top-down flow is not mere attentional modulation. It is the core flow of structured content itself. The forward-flowing signal, by contrast, has now morphed into a stream of residual error. I want to suggest, however, that we treat this apparently radical inversion with some caution. There are two reasons for this—one conceptual, and one empirical.

The first (conceptual) reason for caution is that the “error signal” in a trained-up predictive coding scheme is highly informative. Prediction error signals carry detailed information

about the mismatched content itself. Prediction errors are thus as structured and nuanced in their implications as the model-based predictions relative to which they are computed. This means that, in a very real sense, the prediction error signal is not a mere proxy for incoming sensory information – it *is* sensory information. Thus, suppose you and I play a game in which I (the “higher, predicting level”) try to describe to you (the “lower level”) the scene in front of your eyes. I can’t see the scene directly, but you can. I do, however, believe that you are in some specific room (the living room in my house, say) that I have seen in the past. Recalling that room as best I can, I say to you “there’s a vase of yellow flowers on a table in front of you”. The game then continues like this. If you are silent, I take that as your agreeing to my description. But if I get anything that matters wrong, you must tell me what I got wrong. You might say “the flowers are yellow”. You thus provide an error signal that invites me to try again in a rather specific fashion—that is, to try again with respect to the colour of the flowers in the vase. The next most probable colour, I conjecture, is red. I now describe the scene in the same way but with red flowers. Silence. We have settled into a mutually agreeable description.³

The point to note is that your “error signal” carried some quite specific information. In the pragmatic context of your silence regarding all other matters, the content might be glossed as “there is indeed a vase of flowers on the table in front of me but they are not yellow”. This is a pretty rich message. Indeed, it does not (content-wise) seem different in kind to the down-

ward-flowing predictions themselves. Prediction error signals are thus richly informative, and as such (I would argue) not radically different to sensory information itself. This is unsurprising, since mathematically (as Karl Friston has pointed out⁴) sensory information and prediction error are informationally identical, except that the latter are centred on the predictions. To see this, reflect on the fact that prediction error is just the original information minus the prediction. It follows that the original information is given by the prediction error plus the prediction. Prediction error is simply error relative to some specific prediction and as such it flags the sensory information that is as yet unexplained. The forward flow of prediction error thus constitutes a *forward flow of sensory information relative to specific predictions*.

There is more to the story at this point, since the (complex, non-linear) ways in which downward-flowing predictions interact are importantly different to the (simple, linear) effects of upward-flowing error signals. Non-linearities characterize the multi-level construction of the predictions, which do the “heavy lifting”, while the prediction error signals are free to behave additively (since all the complex webs of linkage are already in place). But the bottom line is that prediction error does not replace sensory information in any mysterious or conceptually challenging fashion, since prediction error is nothing other than that sensory information that has yet to be explained.

The second (empirical) reason for caution is that it is, in any case, only one specific implementation of the predictive brain story depicts the forward-flow as consisting solely of prediction error. An alternative implementation (due to Spratling 2008 and 2010—and see discussion in Spratling 2013) implements the same key principles using a different flow of prediction and error, and described by a variant mathematical framework. This illustrates the urgent need to explore multiple variant architectures for prediction error minimization. In fact, the PP schema occupies just one point in a large and complex space of probabilistic generative-

³ To complete the image using this parlour game, we’d need to add a little more structure to reflect the hierarchical nature of the message-passing scheme. We might thus imagine many even-higher-level “prediction agents” working together to predict which room (house, world, etc.) the layers below are currently responding to. Should sufficient prediction error signals accrue, this ensemble might abandon the hypothesis that signals are coming in from the living room, suggesting instead that they are from the boudoir, or the attic. In this grander version (which recalls the “mixtures of experts” model in machine learning—see Jordan & Jacobs 1994)—there are teams (and teams of teams) of specialist prediction agents, all trying (guided top-down by the other prediction agents, and bottom-up by prediction errors from the level below) to decide which specialists should handle the current sensory barrage. Each higher-level “prediction agent”, in this multi-level version, treats activity at the level below as sensory information, to be explained by the discovery of apt top-down predictions.

⁴ Personal communication.

model-based approaches, and there are many possible architectures and possible ways of combining top-down predictions and bottom-up sensory information in this general vicinity. These include foundational work by Hinton and colleagues on deep belief networks (Hinton & Salakhutdinov 2006; Hinton et al. 2006), work that shares a core emphasis on the use of prediction and probabilistic multi-level generative models, as well as recent work combining connectionist principles with Bayesian angles (see McClelland 2013 and Zorzi et al. 2013). Meanwhile, roboticists such as Tani (2007), Saegusa et al. (2008), Park et al. (2012), Pezzulo (2008), and Mohan et al. (2010) explore the use of a variety of prediction-based learning routines as a means of grounding higher cognitive functions in the solid bedrock of sensorimotor engagements with the world. Only by considering the full space of possible prediction-and-generative-model-based architectures and strategies can we start to ask truly pointed experimental questions about the brain and about biological organisms; questions that might one day favor one of these models (or, more likely, one coherent sub-set of models⁵) over the rest, or else may reveal deep faults and failings among their substantial common foundations.

2.2 Motor control is just more top-down sensory prediction

I shall, however, continue to concentrate upon the specific explanatory schema implied by PP, as this represents (it seems to me) the most comprehensive and neuroscientifically well-grounded vision of the predictive mind currently available. What makes PP especially interesting—and conceptually challenging—is the seamless integration of perception and action achieved under the rubric of “active inference”.

To understand this, consider the motor system. The motor system (like the visual cortex) displays a complex hierarchical structure.⁶

Such a structure allows complex behaviors to be specified, at higher levels, in compact ways, the implications of which can be progressively unpacked at the lower levels. The traditional way of conceptualizing the difference, however, is that in the case of motor control we imagine a downwards flow of information, whereas in the case of the visual cortex we imagine an upwards flow. Descending pathways in the motor cortex, this traditional picture suggests, should correspond functionally to ascending pathways in the visual cortex. This is not, however, the case. Within the motor cortex the downwards connections (descending projections) are “anatomically and physiologically more like backwards connections in the visual cortex than the corresponding forward connections” (Adams et al. 2013, p. 1).

This is suggestive. Where we might have imagined the functional anatomy of a hierarchical motor system to be some kind of inverted image of that of the perceptual system, instead the two seem fundamentally alike.⁷ The explanation, PP suggests, is that the downwards connections, in both cases, take care of essentially the same kind of business—namely the business of predicting sensory stimulation. Predictive processing models subvert, we saw, the traditional picture with respect to perception. In PP, compact higher-level encodings are part of an apparatus that tries to predict the play of energy across the sensory surfaces. The same story applies, recent extensions (see below) of PP suggest, to the motor case. The difference is that motor control is, in a certain sense, subjunctive. It involves predicting the non-actual sensory trajectories that *would* ensue *were* we performing some desired action. Reducing prediction er-

archy is fluid in that the information-flows it supports are reconfigurable moment-by-moment (by, for example, changing β and theta band oscillations —see Bastos et al. 2015). In addition, PP dispenses entirely with the traditional idea (nicely reviewed, and roundly rejected, in Churchland et al. 1994) that earlier levels must complete their tasks before passing information “up” the hierarchy. The upshot is that the PP models are much closer to dynamical systems accounts than to traditional, feed forward, hierarchical ones.

⁷ For the full story, see Adams et al. (2013). In short: “[t]he descending projections from motor cortex share many features with top-down or backward connections in visual cortex; for example, corticospinal projections originate in infragranular layers, are highly divergent and (along with descending cortico-cortical projections) target cells expressing NMDA receptors” (Adams et al. 2013, p. 1).

⁵ One such subset is, of course, the set of hierarchical dynamic models (see Friston 2008).

⁶ The appeal to hierarchical structure in PP, it should be noted, is substantially different to that familiar from treatments such as Felleman & Van Essen (1991). Although I cannot argue for this here (for more on this see Clark 2013b; *in press*) the PP hier-

rors calculated against these non-actual states then serves (in ways we are about to explore) to make them actual. We predict the sensory consequences of our own action and this brings the actions about.

The upshot is that the downwards connections, in both the motor and the sensory cortex, carry complex predictions, and the upwards connections carry prediction errors. This explains the otherwise “paradoxical” (Shipp et al. 2013, p. 1) fact that the functional circuitry of the motor cortex does not seem to be inverted with respect to that of the sensory cortex. Instead, the very distinction between the motor and the sensory cortex is now eroded—both are in the business of top-down prediction, though the kind of thing they predict is (of course) different. The motor cortex here emerges, ultimately, as a multimodal sensorimotor area issuing predictions in both proprioceptive and other modalities.

In this way, PP models have been extended (under the umbrella of “active inference”—see Friston 2009; Friston et al. 2011) to include the control of action. This is accomplished by predicting the flow of sensation (especially that of proprioception) that would occur were some target action to be performed. The resulting cascade of prediction error is then quashed by moving the bodily plant so as to bring the action about. Action thus results from our own predictions concerning the flow of sensation—a version of the “ideomotor” theory of James (1890) and Lotze (1852), according to which the very idea of moving, when unimpeded by other factors, is what brings the moving about. The resulting schema is one in which:

The perceptual and motor systems should not be regarded as separate but instead as a single active inference machine that tries to predict its sensory input in all domains: visual, auditory, somatosensory, interoceptive and, in the case of the motor system, proprioceptive. (Adams et al. 2013, p. 4)

In the case of motor behaviors, the key driving predictions, Friston and colleagues suggest, are

predictions of the proprioceptive patterns⁸ that would ensue were the action to be performed (see Friston et al. 2010). To make an action come about, the motor plant responds so as to cancel out proprioceptive prediction errors. In this way, predictions of the unfolding proprioceptive patterns that would be associated with the performance of some action serve to bring that action about. Proprioceptive predictions directly elicit motor actions (so traditional motor commands are simply replaced by those proprioceptive predictions).

This erases any fundamental computational line between perception and the control of action. There remains, to be sure, an obvious (and important) difference in direction of fit. Perception here matches neural hypotheses to sensory inputs, and involves “predicting the present”; while action brings unfolding proprioceptive inputs into line with neural predictions. The difference, as Elizabeth Anscombe (1957) famously remarked,⁹ is akin to that between consulting a shopping list to select which items to purchase (thus letting the list determine the contents of the shopping basket) and listing some actually purchased items (thus letting the contents of the shopping basket determine the list). But despite this difference in direction of fit, the underlying form of the neural computations is now revealed to be the same. Indeed, the main difference between the motor and the visual cortex, on this account, lies more in what kind of thing (for example, the proprioceptive consequences of a trajectory of motion) is predicted, rather than in how it is predicted. The upshot is that:

The primary motor cortex is no more or less a motor cortical area than striate (visual) cortex. The only difference

⁸ Proprioception is the “inner” sense that informs us about the relative locations of our bodily parts and the forces and efforts that are being applied. It is to be distinguished from exteroceptive (i.e., standard perceptual) channels such as vision and audition, and from interoceptive channels informing us of hunger, thirst, and states of the viscera. Predictions concerning the latter may (see Seth 2013 and Pezzulo 2014) play a large role in the construction of feelings and emotions.

⁹ Anscombe’s target was the distinction between desire and belief, but her observations about direction of fit generalize (as Shea 2013 notes) to the case of actions, here conceived as the motoric outcomes of certain forms of desire.

between the motor cortex and visual cortex is that one predicts retinotopic input while the other predicts proprioceptive input from the motor plant. (Friston et al. 2011, p. 138)

Perception and action here follow the same basic logic and are implemented using the same computational strategy. In each case, the systemic imperative remains the same: the reduction of ongoing prediction error. This view has two rather radical consequences, to which we shall now turn.

2.3 Efference copies and distinct “controllers” are replaced by top-down predictions

A long tradition in the study of motor control invokes a “forward model” of the likely sensory consequences of our own motor commands. In this work, a copy of the motor command (known as the “efference copy”; Von Holst 1954) is processed using the forward model. This model captures (or “emulates”—see Grush 2004) the relevant biodynamics of the motor plant, enabling (for example) a rapid prediction of the likely feedback from the sensory peripheries. It does this by encoding the relationship between motor commands and predicted sensory outcomes. The motor command is thus captured using the efference copy which, fed to the forward model, yields a prediction of the sensory outcome (this is sometimes called the “corollary discharge”). Comparisons between the actual and the predicted sensory input are thus enabled.

But motor control, in the leading versions of this kind of account, requires in addition the development and use of a so-called “inverse model” (see e.g., Kawato 1999; Franklin & Wolpert 2011). Where the forward model maps current motor commands in order to predicted sensory effects, the inverse model (also known as a controller) “performs the opposite transformation [...] determining the motor command required to achieve some desired outcome” (Wolpert et al. 2003, p. 595). Learning and deploying an inverse model appropriate to some

task is, however, generally much more demanding than learning the forward model, and requires solving a complex mapping problem (linking the desired end-state to a nested cascade of non-linearly interacting motor commands) while effecting transformations between varying co-ordinate schemes (e.g., visual to muscular or proprioceptive—see e.g., Wolpert et al. 2003, pp. 594–596).

PP (the full “action-inclusive” version just described) shares many key insights with this work. They have common a core emphasis on the prediction-based learning of a forward (generative) model, which is able to anticipate the sensory consequences of action. But active inference, as defended by Friston and others—see e.g., Friston (2011); Friston et al. (2012)—dispenses with the inverse model or controller, and along with it the need for efference copy of the motor command. To see how this works, consider that action is here reconceived as a direct consequence of predictions (spanning multiple temporal and spatial scales) about trajectories of motion. Of special importance here are predictions about proprioceptive consequences that implicitly minimize various energetic costs. Subject to the full cascade of hierarchical top-down processing, a simple motor command now unfolds into a complex set of predictions concerning both proprioceptive and exteroceptive effects. The proprioceptive predictions then drive behavior, causing us to sample the world in the ways that the current winning hypothesis dictates.¹⁰

Such predictions can be couched, at the higher levels, in terms of desired states or trajectories specified using extrinsic (world-centered, limb-centered) co-ordinates. This is possible because the required translation into intrinsic (muscle-based) co-ordinates is then devolved to what are essentially classical reflex arcs set up to quash proprioceptive prediction errors. Thus:

if motor neurons are wired to suppress proprioceptive prediction errors in the dorsal horn of the spinal cord, they effect-

¹⁰ For a simulation-based demonstration of the overall shape of the PP account, see Friston et al. (2012). These simulations, as the authors note, turn out to implement the kind of “active vision” account put forward in Wurtz et al. (2011).

ively implement an inverse model, mapping from desired sensory consequences to causes in intrinsic (muscle-based) coordinates. In this simplification of conventional schemes, descending motor commands become topdown predictions of proprioceptive sensations conveyed by primary and secondary sensory afferents. (Friston 2011, p. 491)

The need (prominent in approaches such as Kawato 1999; Wolpert et al. 2003; and Franklin & Wolpert 2011) for a distinct inverse model/optimal control calculation has now disappeared. In its place we find a more complex forward model mapping prior beliefs about desired trajectories to sensory consequences, some of which (the “bottom level” proprioceptive ones) are automatically fulfilled.

The need for efference copy has also disappeared. This is because descending signals are already (just as in the perceptual case) in the business of predicting sensory (both proprioceptive and exteroceptive) consequences. By contrast, so-called “corollary discharge” (encoding predicted sensory outcomes) is now endemic and pervades the downwards cascade, since:

[...] every backward connection in the brain (that conveys topdown predictions) can be regarded as corollary discharge, reporting the predictions of some sensorimotor construct. (Friston 2011, p. 492)

This proposal may, on first encounter, strike the reader as quite implausible and indeed too radical. Isn’t an account of the functional significance and neurophysiological reality of efference copy one of the major success stories of contemporary cognitive and computational neuroscience? In fact, most (perhaps all) of the evidence often assumed to favour that account is, on closer examination, simply evidence of the pervasive and crucial role of forward models and corollary discharge—it is evidence, that is to say, for just those parts of the traditional story that are preserved (and made even more central) by PP. For example, Sommer & (Wurtz’s influential (2008) review paper makes very little

mention of efference copy as such, but makes widespread use of the more general concept of corollary discharge—though as those authors note, the two terms are often used interchangeably in the literature. A more recent paper, Wurtz et al. (2011), mentions efference copy only once, and does so only to merge it with discussions of corollary discharge (which then occur 114 times in the text). Similarly, there is ample reason to believe that the cerebellum plays a special role here, and that that role involves making or optimizing perceptual predictions about upcoming sensory events (Bastian 2006; Roth et al. 2013). But such a role is, of course, entirely consistent with the PP picture. This shows, I suggest, that it is the general concept of forward models (as used by e.g., Miall & Wolpert 1996) and corollary discharge, rather than the more specific concept of efference copy as we defined it above, that enjoys the clearest support from both experimental and cognitive neuroscience.

Efference copy figures prominently, of course, in one particular set of computational proposals. These proposals concern (in essence) the positioning of forward models and corollary discharges within a putative larger cognitive architecture involving multiple paired forward and inverse models. In these “paired forward inverse model” architectures (see e.g., Wolpert & Kawato 1998; Haruno et al. 2003) motor commands are copied to a stack of separate forward models that are used to predict the sensory consequences of actions. But acquiring and deploying such an architecture, as even its strongest advocates concede, poses a variety of extremely hard computational challenges (see Franklin & Wolpert 2011). The PP alternative neatly sidesteps many of these problems—as we shall see in section 2.4. The heavy lifting that is usually done by traditional efference copy, inverse models, and optimal controllers is now shifted to the acquisition and use of the predictive (generative) model—i.e., the right set of prior probabilistic “beliefs”. This is potentially advantageous if (but only if) we can reasonably assume that these beliefs “emerge naturally as top-down or empirical priors during hierarchical perceptual inference” (Friston 2011, p. 492).

The deeper reason that efference copy may be said to have disappeared in PP is thus that the whole (problematic) structure of paired forward and inverse models is absent. It is not needed, because some of the predicted sensory consequences (the predicted proprioceptive trajectories) act as motor commands already. As a result, there are no distinct motor commands to copy, and (obviously) no efference copies as such. But one could equally well describe the forward-model-based predictions of proprioceptive trajectories as “minimal motor commands”: motor commands that operate (in essence) by specifying results rather than by exerting fine-grained limb and joint control. These minimal motor commands (proprioceptive predictions) clearly influence the even wider range of predictions concerning the exteroceptive sensory consequences of upcoming actions. The core functionality that is normally attributed to the action of efference copy is thus preserved in PP, as is the forward-model-based explanation of core phenomena, such as the finessing of time-delays (Bastian 2006) and the stability of the visual world despite eye-movements (Sommer & Wurtz 2006; 2008).

2.4 Cost functions are absorbed by predictions.

Active inference also sidesteps the need for explicit cost or value functions as a means of selecting and sculpting motor response. It does this (Friston 2011; Friston et al. 2012) by, in essence, building these in to the generative model whose probabilistic predictions combine with sensory inputs in order to yield behaviors. Simple examples of cost or value functions (that might be applied to sculpt and select motor behaviors) include minimizing “jerk” (the rate of change of acceleration of a limb during some behavior) and minimizing rate of change of torque (for these examples see Flash & Hogan 1985 and Uno et al. 1989 respectively). Recent work on “optimal feedback control” minimizes more complex “mixed cost functions” that address not just bodily dynamics but also systemic noise and the required accuracy of outcomes (see Todorov 2004; Todorov & Jordan 2002).

Such cost functions (as Friston 2011, p. 496 observes) resolve the many-one mapping problem that afflicts classical approaches to motor control. There are many ways of using one’s body to achieve a certain goal, but the action system has to choose one way from the many available. Such devices are not, however, needed within the framework on offer, since:

In active inference, these problems are resolved by prior beliefs about the trajectory (that may include minimal jerk) that uniquely determine the (intrinsic) consequences of (extrinsic) movements. (Friston 2011, p. 496)

Simple cost functions are thus folded into the expectations that determine trajectories of motion. But the story does not stop there. For the very same strategy applies to the notion of desired consequences and rewards at all levels. Thus we read that:

Crucially, active inference does not invoke any “desired consequences”. It rests only on experience-dependent learning and inference: experience induces prior expectations, which guide perceptual inference and action. (Friston et al. 2011, p. 157)

Notice that there is no *overall* computational advantage to be gained by this reallocation of duties. Indeed, Friston himself is clear that:

[...] there is no free lunch when replacing cost functions with prior beliefs [since] it is well-known [Littman et al. (2001)] that the computational complexity of a problem is not reduced when formulating it as an inference problem. (2011, p. 492)

Nonetheless, it may well be that this reallocation (in which cost functions are treated as priors) has conceptually and strategically important consequences. It is easy, for example, to specify whole paths or trajectories using prior beliefs about (you guessed it) paths and trajectories! Scalar reward functions, by contrast, specify points or peaks. The upshot is that everything

that can be specified by a cost function can be specified by some prior over trajectories, but not vice versa.

Related concerns have led many working roboticists to argue that explicit cost-function-based solutions are inflexible and biologically unrealistic, and should be replaced by approaches that entrain actions in ways that implicitly exploit the complex attractor dynamics of embodied agents (see e.g., [Thelen & Smith 1994](#); [Mohan & Morasso 2011](#); [Feldman 2009](#)). One way to imagine this broad class of solutions (for a longer discussion, see [Clark 2008](#), Ch. 1) is by thinking of the way you might control a wooden marionette simply by moving the strings attached to specific body parts. In such cases:

The distribution of motion among the joints is the “passive” consequence of the [...] forces applied to the end-effectors and the “compliance” of different joints. ([Mohan & Morasso 2011](#), p. 5)

Solutions such as these, which make maximal use of learnt or inbuilt “synergies” and the complex bio-mechanics of the bodily plant, can be very fluently implemented (see [Friston 2011](#); [Yamashita & Tani 2008](#)) using the resources of active inference and (attractor-based) generative models. For example, [Namikawa et al. \(2011\)](#) show how a generative model with multi-timescale dynamics enables a fluent and decomposable (see also [Namikawa & Tani 2010](#)) set of motor behaviors. In these simulations:

Action per se, was a result of movements that conformed to the proprioceptive predictions of [...] joint angles [and] perception and action were both trying to minimize prediction errors throughout the hierarchy, where movement minimized the prediction errors at the level of proprioceptive sensations. ([Namikawa et al. 2011](#), p. 4)

Another example (which we briefly encountered in the previous section) is the use of downward-flowing prediction to side-step the need to transform desired movement trajectories from

extrinsic (task-centered) to intrinsic (e.g., muscle-centered) co-ordinates: an “inverse problem” that is said to be both complex and ill-posed ([Feldman 2009](#); [Adams et al. 2013](#), p. 8). In active inference the prior beliefs that guide motor action already map predictions couched (at high levels) in extrinsic frames of reference onto proprioceptive effects defined over muscles and effectors, simply as part and parcel of ordinary online control.

By re-conceiving cost functions as implicit in bodies of expectations concerning trajectories of motion, PP-style solutions sidestep the need to solve difficult (often intractable) optimality equations during online processing (see [Friston 2011](#); [Mohan & Morasso 2011](#)) and—courtesy of the complex generative model—fluidly accommodate signaling delays, sensory noise, and the many-one mapping between goals and motor programs. Alternatives requiring the distinct and explicit computation of costs and values thus arguably make unrealistic demands on online processing, fail to exploit the helpful characteristics of the physical system, and lack biologically plausible means of implementation.

These various advantages come, however, at a price. For the full PP story now shifts much of the burden onto the acquisition of those prior “beliefs”—the multi-level, multi-modal webs of probabilistic expectation that together drive perception and action. This may turn out to be a better trade than it at first appears, since (see [Clark in press](#)) PP describes a biologically plausible architecture that is just about maximally well-suited to installing the requisite suites of prediction, through embodied interactions with the training environments that we encounter, perturb, and—at several slower timescales—actively construct.

3 Putting predictive processing, body, and world together again

An important feature of the full PP account (see [Friston 2009](#); [Hohwy 2013](#); [Clark in press](#)) is that the impact of specific prediction error signals can be systematically varied according to their estimated certainty or “precision”. The precision of a specific prediction error is

its inverse variance—the size (if you like) of its error bars. Precision estimation thus has a kind of meta-representational feel, since we are, in effect, estimating the uncertainty of our own representations of the world. These ongoing (task and context-varying) estimates alter the weighting (the gain or volume, to use the standard auditory analogy) on select prediction error units, so as to increase the impact of task-relevant, reliable information. One key effect of this is to allow the brain to vary the balance between sensory inputs and prior expectations at different levels (see [Friston 2009](#), p. 299) in ways sensitive to task and context.¹¹ High-precision prediction errors have greater gain, and thus play a larger role in driving processing and response. More generally, variable precision-weighting may be seen as the PP mechanism for implementing a wide range of attentional effects (see [Feldman & Friston 2010](#)).

Subtle applications of this strategy, as we shall shortly see, allow PP to nest simple (“quick and dirty”) solutions within the larger context of a fluid, re-configurable inner economy; an economy in which rich, knowledge-based strategies and fast, frugal solutions are now merely different expressions of a unified underlying web of processing. Within that web, changing ensembles of inner resources are repeatedly recruited, forming and dissolving in ways determined by external context, current needs, and (importantly) by flexible precision-weighting reflecting ongoing estimations of our own uncertainty. This process of inner recruitment is itself constantly modulated, courtesy of the complex circular causal dance of sensorimotor engagement, by the evolving state of the external environment. In this way (as I shall now argue) many key insights from work on embodiment and situated, world-exploiting action may be comfortably accommodated within the emerging PP framework.

¹¹ Malfunctions of this precision-weighting apparatus have recently been implicated in a number of fascinating proposals concerning the origins and persistence of various forms of mental disturbance, including the emergence of delusions and hallucinations in schizophrenia, “functional motor and sensory symptoms”, Parkinson’s disease, and autism—see [Fletcher & Frith \(2009\)](#), [Frith & Friston \(2012\)](#), [Adams et al. \(2012\)](#), [Brown et al. \(2013\)](#), [Edwards et al. \(2012\)](#), and [Pellicano & Burr \(2012\)](#).

3.1 Nesting simplicity within complexity

Consider the well-known “outfielder’s problem”: running to catch a fly ball in baseball. Giving perception its standard role, we might assume that the job of the visual system is to transduce information about the current position of the ball so as to allow a distinct “reasoning system” to project its future trajectory. Nature, however, seems to have found a more elegant and efficient solution. The solution, a version of which was first proposed in [Chapman \(1968\)](#), involves running in a way that seems to keep the ball moving at a constant speed through the visual field. As long as the fielder’s own movements cancel any apparent changes in the ball’s optical acceleration, she will end up in the location where the ball hits the ground. This solution, OAC (Optical Acceleration Cancellation), explains why fielders, when asked to stand still and simply predict where the ball will land, typically do rather badly. They are unable to predict the landing spot because OAC is a strategy that works by means of moment-by-moment self-corrections that, crucially, involve the agent’s own movements. The suggestion that we rely on such a strategy is also confirmed by some interesting virtual reality experiments in which the ball’s trajectory is suddenly altered in flight, in ways that could not happen in the real world—see [Fink et al. 2009](#)). OAC is a succinct case of fast, economical problem-solving. The canny use of data available in the optic flow enables the catcher to sidestep the need to deploy a rich inner model to calculate the forward trajectory of the ball.¹²

Such strategies are suggestive (see also [Maturana & Varela 1980](#)) of a very different role of the perceptual coupling itself. Instead of using sensing to get enough information inside, past the visual bottleneck, so as to allow the reasoning system to “throw away the world” and solve the problem wholly internally, such strategies use the sensor as *an open conduit allowing environmental magnitudes to exert a constant influence on behavior*. Sensing is here

¹² There are related accounts of how dogs catch Frisbees—a rather more demanding task due to occasional dramatic fluctuations in the flight path (see [Shaffer et al. 2004](#)).

depicted as the opening of a channel, with successful whole-system behavior emerging when activity in this channel is kept within a certain range. In such cases:

[T]he focus shifts from accurately representing an environment to continuously engaging that environment with a body so as to stabilize appropriate co-ordinated patterns of behaviour. (Beer 2000, p. 97)

These focal shifts may be fluidly accommodated within the PP framework. To see how, recall that “precision weighting” alters the gain on specific prediction error units, and thus provides a means of systematically varying the relative influence of different neural populations. The most familiar role of such manipulations is to vary the balance of influence between bottom-up sensory information and top-down model-based expectation. But another important role is the implementation of fluid and flexible forms of large-scale “gating” among neural populations. This works because very low-precision prediction errors will have little or no influence upon ongoing processing, and will fail to recruit or nuance higher-level representations. Altering the distribution of precision weightings thus amounts, as we saw above, to altering the “simplest circuit diagram” (Aertsen & Preißl 1991) for current processing. When combined with the complex, cascading forms of influence made available by the apparatus of top-down prediction, the result is an inner processing economy that is (see Clark in press) “maximally context-sensitive”.

This suggests a new angle upon the outfielder’s problem. Here too, already-active neural predictions and simple, rapidly-processed perceptual cues must work together (if PP is correct) to determine a pattern of precision-weightings for different prediction-error signals. This creates a pattern of effective connectivity (a temporary distributed circuit) and, within that circuit, it sets the balance between top-down and bottom-up modes of influence. In the case at hand, however, efficiency demands selecting a circuit in which visual sensing is used to cancel the optical acceleration of the fly ball.

This means giving high weighting to the prediction errors associated with cancelling the vertical acceleration of the ball’s optical projection, and (to put it bluntly) not caring very much about anything else. Apt precision weightings here function to select *what to predict* at any given moment. They may thus select a pre-learned, fast, low-cost strategy for solving a problem, as task and context dictate. Contextually-recruited patterns of precision weighting thus accomplish a form of set-selection or strategy switching—an effect already demonstrated in some simple simulations of cued reaching under the influence of changing tonic levels of dopamine firing—see Friston et al. (2012).

Fast, efficient solutions have also been proposed in the context of reasoning and choice. In an extensive literature concerning choice and decision-making, it has been common to distinguish between “model-based” and “model-free” approaches (see e.g., Dayan & Daw 2008; Dayan 2012; Wolpert et al. 2003). Model-based strategies rely, as their name suggests, on a model of the domain that includes information about how various states (worldly situations) are connected, thus allowing a kind of principled estimation (given some cost function) of the value of a putative action. Such approaches involve the acquisition and the (computationally challenging) deployment of fairly rich bodies of information concerning the structure of the task-domain. Model-free strategies, by contrast, are said to “learn action values directly, by trial and error, without building an explicit model of the environment, and thus retain no explicit estimate of the probabilities that govern state transitions” (Gläscher et al. 2010, p. 585). Such approaches implement “policies” that typically exploit simple cues and regularities while nonetheless delivering fluent, often rapid, response.

The model-based/model-free distinction is intuitive, and resonates with old (but increasingly discredited) dichotomies between reason and habit, and between analytic evaluation and emotion. But it seems likely that the image of parallel, functionally independent, neural subsystems will not stand the test of time. For example, a recent functional Magnetic Resonance Imaging (fMRI) study (Daw et al. 2011) sug-

gests that rather than thinking in terms of distinct (functionally isolated) model-based and model-free learning systems, we may need to posit a single “more integrated computational architecture” (Daw et al. 2011, p. 1204), in which the different brain areas most commonly associated with model-based and model-free learning (pre-frontal cortex and dorsolateral striatum, respectively) *each* trade in both model-free and model-based modes of evaluations and do so “in proportions matching those that determine choice behavior” (Daw et al. 2011, p. 1209). Top-down information, (Daw et al. (2011) suggest, might then control the way different strategies are combined in differing contexts for action and choice. Within the PP framework, this would follow from the embedding of shallow “model-free” responses within a deeper hierarchical generative model. By thus combining the two modes within an overarching model-based economy, inferential machinery can, by and large, identify the appropriate contexts in which to deploy the model-free (“habitual”) schemes. “Model-based” and “model-free” modes of valuation and response, if this is correct, name extremes along a single continuum, and may appear in many mixtures and combinations determined by the task at hand.

This suggests a possible reworking of the popular suggestion (Kahneman 2011) that human reasoning involves the operation of two functionally distinct systems: one for fast, automatic, “habitual” response, and the other dedicated to slow, effortful, deliberative reasoning. Instead of a truly dichotomous inner organization, we may benefit from a richer form of organization in which fast, habitual, or heuristically-based modes of response are often the default, but within which a large variety of possible strategies may be available. Humans and other animals would thus deploy multiple—rich, frugal and all points in between—strategies defined across a fundamentally unified web of neural resources (for some preliminary exploration of this kind of more integrated space, see Pezzulo et al. 2013). Some of those strategies will involve the canny use of environmental structure – efficient embodied prediction machines, that is to say, will often deploy minimal

neural models that benefit from repeated calls to world-altering action (as when we use a few taps of the smartphone to carry out a complex calculation).

Nor, finally, is there any fixed limit to the complexities of the possible strategic embeddings that might occur even within a single more integrated system. We might, for example, use some quick-and-dirty heuristic strategy to identify a context in which to use a richer one, or use intensive model-exploring strategies to identify a context in which a simpler one will do. From this emerging vantage point the very distinction between model-based and model-free response (and indeed between System 1 and System 2) looks increasingly shallow. These are now just convenient labels for different admixtures of resource and influence, each of which is recruited in the same general way as circumstances dictate.¹³

3.2 Being human

There is nothing specifically human, however, about the suite of mechanisms explored above. The basic elements of the predictive processing story, as Roepstorff (2013, p. 45) correctly notes, may be found in many types of organism and model-system. The neocortex (the layered structure housing cortical columns that provides the most compelling neural implementation for predictive processing machinery) displays some dramatic variations in size but is common to all mammals. What, then, makes us (superficially at least) so very different? What is it that allows us—unlike dogs, chimps, or dolphins—to latch on to distal hidden causes that include not just food, mates, and relative social rankings, but also neurons, predictive processing, Higgs bosons, and black holes?

One possibility (Conway & Christiansen 2001) is that adaptations of the human neural apparatus have somehow conspired to create, in us, an even more complex and context-flexible

¹³ Current thinking about switching between model-free and model-based strategies places them squarely in the context of hierarchical inference, through the use of “Bayesian parameter averaging”. This essentially associates model-free schemes with simpler (less complex) lower levels of the hierarchy that may, at times, need to be contextualized by (more complex) higher levels.

hierarchical learning system than is found in other animals. Insofar as the predictive processing framework allows for rampant context-dependent influence within the distributed hierarchy, the same basic operating principles might (given a few new opportunities for routing and influence) result in the emergence of qualitatively novel forms of behavior and control. Such changes might explain why human agents display what [Spivey \(2007, p. 169\)](#) describes as an “exceptional sensitivity to hierarchical structure in *any* time-dependent signal”.

Another (possibly linked, and certainly highly complementary) possibility involves a potent complex of features of human life, in particular our ability to engage in temporally coordinated social interaction (see [Roepstorff et al. 2010](#)) and our ability to construct artifacts and design environments. Some of these ingredients have emerged in other species too. But in the human case the whole mosaic comes together under the influence of flexible and structured symbolic language (this was the target of the Conway and Christiansen paper mentioned above) and an almost obsessive drive ([Tommasello et al. 2005](#)) to engage in shared cultural practices. We are thus able to redeploy our core cognitive skills in the transformative context of exposure to what [Roepstorff et al. \(2010\)](#) call “patterned sociocultural practices”. These include the use of symbolic codes (encountered as “material symbols” ([Clark 2006](#)) and complex social routines ([Hutchins 1995, 2014](#))—and more general, all the various plays and strategies known as “cognitive niche construction” (see [Clark 2008](#)).

A simple example is the way that learning to perform mental arithmetic has been scaffolded, in some cultures, by the deliberate use of an abacus. Experience with patterns thus made available helps to install appreciation of many complex arithmetical operations and relations (for discussion of this, see [Stigler 1984](#)). The specific example does not matter very much, to be sure, but the general strategy does. In such cases, we structure (and repeatedly re-structure) our physical and social environments in ways that make available new knowledge and skills—see [Landy & Goldstone \(2005\)](#). Prediction-hungry brains, ex-

posed in the course of embodied action to novel patterns of sensory stimulation, may thus acquire forms of knowledge that were genuinely out-of-reach prior to such physical-manipulation-based re-tuning of the generative model. Action and perception thus work together to reduce prediction error against the more slowly evolving backdrop of a culturally distributed process that spawns a succession of designed environments whose impact on the development (e.g., [Smith & Gasser 2005](#)) and unfolding ([Hutchins 2014](#)) of human thought and reason can hardly be overestimated.

To further appreciate the power and scope of such re-shaping, recall that the predictive brain is not doomed to deploy high-cost, model-rich strategies moment-by-moment in a demanding and time-pressured world. Instead, that very same apparatus supports the learning and contextually-determined deployment of low-cost strategies that make the most of body, world, and action. A maximally simple example is painting white lines along the edges of a winding cliff-top road. Such environmental alterations allow the driver to solve the complex problem of keeping the car on the road by (in part) predicting the ebb and flow of various simpler optical features and cues (see e.g., [Land 2001](#)). In such cases, we are building a better world in which to predict, while simultaneously structuring the world to cue the low-cost strategy at the right time.

3.3 Extending the predictive mind

All this suggests a very natural model of “extended cognition” ([Clark & Chalmers 1998; Clark 2008](#)), where this is simply the idea that bio-external structures and operations may sometimes form integral parts of an agent’s cognitive routines. Nothing in the PP framework materially alters, as far as I can tell, the arguments previously presented, both pro and con, regarding the possibility and actuality of genuinely extended cognitive systems.¹⁴ What PP

¹⁴ For a thorough rehearsal of the positive arguments, see [Clark \(2008\)](#). For critiques, see [Rupert \(2004, 2009\)](#), [Adams & Aizawa \(2001\)](#), and [Adams & Aizawa \(2008\)](#). For a rich sampling of the ongoing debate, see the essays in [Menary \(2010\)](#) and [Estany & Sturm \(2014\)](#).

does offer, however, is a specific and highly “extension-friendly” proposal concerning the shape of the specifically neural contribution to cognitive success. To see this, reflect on the fact that known external (e.g., environmental) operations provide—by partly constituting—additional strategies apt for the kind of “meta-model-based” selection described above. This is because actions that engage and exploit specific external resources will now be selected in just the same manner as the inner coalitions of neural resources themselves. Minimal internal models that involve calls to world-recruiting actions may thus be selected in the same way as a purely internal model. The availability of such strategies (of trading inner complexity against real-world action) is the hallmark of embodied prediction machines.

As a simple illustration, consider the work undertaken by [Pezzulo et al. \(2013\)](#). Here, a so-called “Mixed Instrumental Controller” determines whether to choose an action based upon a set of simple, pre-computed (“cached”) values, or by running a mental simulation enabling a more flexible, model-based assessment of the desirability, or otherwise, of actually performing the action. The mixed controller computes the “value of information”, selecting the more informative (but costly) model-based option only when that value is sufficiently high. Mental simulation, in such cases, then produces new reward expectancies that can determine current action by updating the values used to determine choice. We can think of this as a mechanism that, moment-by-moment, determines (as discussed in previous sections) whether to exploit simple, already-cached routines or to explore a richer set of possibilities using some form of mental simulation. It is easy to imagine a version of the mixed controller that determines (on the basis of past experience) the value of the information that it believes would be made available by some kind of cognitive extension, such as the manipulation of an abacus, an iPhone, or a physical model. Deciding when to rest, content with a simple cached strategy, when to deploy a more costly mental simulation, and when to exploit the environment itself as a cognitive resource are thus all options apt for the same

kind of “meta-Bayesian” model-based resolution.

Seen from this perspective, the selection of task-specific inner *neural* coalitions within an interaction-dominated PP economy is entirely on a par with the selection of task-specific *neural-bodily-worldly* ensembles. The recruitment and use of extended (brain-body-world) problem-solving ensembles now turns out to obey many of the same basic rules, and reflects many of the same basic normative principles (balancing efficacy and efficiency, and reflecting complex precision estimations) as does the recruitment of temporary inner coalitions bound by effective connectivity. In each case, what is selected is a temporary problem-solving ensemble (a “temporary task-specific device”—see [Anderson et al. 2012](#)) recruited as a function of context-varying estimations of uncertainty.

4 Conclusion: Towards a mature science of the embodied mind

By self-organizing around prediction error, and by learning a generative rather than a merely discriminative (i.e., pattern-classifying) model, these approaches realize many of the goals of previous work in artificial neural networks, robotics, dynamical systems theory, and classical cognitive science. They self-organize around prediction error signals, perform unsupervised learning using a multi-level architecture, and acquire a satisfying grip—courtesy of the problem decompositions enabled by their hierarchical form—upon structural relations within a domain. They do this, moreover, in ways that are firmly grounded in the patterns of sensorimotor experience that structure learning, using continuous, non-linguaform, inner encodings (probability density functions and probabilistic inference). Precision-based restructuring of patterns of effective connectivity then allow us to nest simplicity within complexity, and to make as much (or as little) use of body and world as task and context dictate.

This is encouraging. It might even be that models in this broad ballpark offer us a first glimpse of the shape of a fundamental and unified science of the embodied mind.

Acknowledgements

This work was supported in part by the AHRC-funded ‘Extended Knowledge’ project, based at the Eidyn research centre, University of Edinburgh.

References

- Adams, F. & Aizawa, K. (2001). The bounds of cognition. *Philosophical Psychology*, 14 (1), 43-64. [10.1080/09515080120033571](https://doi.org/10.1080/09515080120033571)
- (2008). *The bounds of cognition*. Malden, MA: Blackwell Publishing.
- Adams, R. A., Perrinet, L. U. & Friston, K. (2012). Smooth pursuit and visual occlusion: Active inference and oculomotor control in schizophrenia. *PLoS One*, 7 (10), e47502. [10.1371/journal.pone.0047502](https://doi.org/10.1371/journal.pone.0047502)
- Adams, R. A., Shipp, S. & Friston, K. J. (2013). Predictions not commands: Active inference in the motor system. *Brain Structure and Function*, 218 (3), 611-643. [10.1007/s00429-012-0475-5](https://doi.org/10.1007/s00429-012-0475-5)
- Aertsen, A. & Preißl, H. (1991). Dynamics of activity and connectivity in physiological neuronal networks. In H. G. Schuster (Ed.) *Nonlinear dynamics and neuronal networks* (pp. 281-302). Weinheim, GER: VCH Verlag.
- Anderson, M. L., Richardson, M. & Chemero, A. (2012). Eroding the boundaries of cognition: Implications of embodiment. *Topics in Cognitive Science*, 4 (4), 717-730. [10.1111/j.1756-8765.2012.01211.x](https://doi.org/10.1111/j.1756-8765.2012.01211.x)
- Anscombe, G. E. M. (1957). *Intention*. Oxford, UK: Basil Blackwell.
- Ballard, D. (1991). Animate vision. *Artificial Intelligence*, 48, 57-86. [10.1016/0004-3702\(91\)90080-4](https://doi.org/10.1016/0004-3702(91)90080-4)
- Ballard, D., Hayhoe, M., Pook, P. & Rao, R. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, 20 (4), 723-767.
- Bastian, A. (2006). Learning to predict the future: The cerebellum adapts feedforward movement control. *Current opinion in neurobiology*, 16 (6), 645-649.
- Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P. & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, 76 (4), 695-711. [10.1016/j.neuron.2012.10.038](https://doi.org/10.1016/j.neuron.2012.10.038)
- Bastos, A. M., Vezoli, J., Bosman, C. A., Schoffelen, J.-M., Oostenveld, R., Dowdall, J. R., De Weerd, P., Kennedy, H. & Fries, P. (2015). Visual areas exert feedforward and feedback influences through distinct frequency channels. *Neuron*. [10.1016/j.neuron.2014.12.018](https://doi.org/10.1016/j.neuron.2014.12.018)
- Beer, R. D. (2000). Dynamical approaches to cognitive science. *Trends in Cognitive Sciences*, 4 (3), 91-99. [10.1016/S1364-6613\(99\)01440-0](https://doi.org/10.1016/S1364-6613(99)01440-0)
- Brooks, R. (1991). Intelligence without representation. *Artificial Intelligence*, 47, 139-159. [10.1.1.12.1680](https://doi.org/10.1.1.12.1680)
- Brown, H., Adams, R. A., Parees, I., Edwards, M. & Friston, K. (2013). Active inference, sensory attenuation

- and illusions. *Cognitive Processing*, 14 (4), 411-427. [10.1007/s10339-013-0571-3](#)
- Chapman, S. (1968). Catching a baseball. *American Journal of Physics*, 36, 868-870.
- Churchland, P. S., Ramachandran, V. S. & Sejnowski, T. J. (1994). A critique of pure vision. In C. Koch & J. L. Davis (Eds.) *Large Scale Neuronal Theories of the Brain* (pp. 23-60). Cambridge, MA: MIT Press.
- Clark, A. (2006). Language, embodiment and the cognitive niche. *Trends in Cognitive Sciences*, 10 (8), 370-374. [10.1016/j.tics.2006.06.012](#)
- (2008). *Supersizing the mind: Action, embodiment, and cognitive extension*. New York, NY: Oxford University Press.
- (2013a). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36 (3), 181-204. [10.1017/S0140525X12000477](#)
- (2013b). The many faces of precision. *Frontiers in Theoretical and Philosophical Psychology*, 4 (270), 1-9. [10.3389/fpsyg.2013.00270](#)
- (in press). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. New York, NY: Oxford University Press.
- Clark, A. & Chalmers, D. (1998). The extended mind. *Analysis*, 58 (1), 7-19. [10.1111/1467-8284.00096](#)
- Conway, C. & Christiansen, M. (2001). Sequential learning in non-human primates. *Trends in Cognitive Sciences*, 5 (12), 539-546. [10.1016/S1364-6613\(00\)01800-3](#)
- Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69, 1204-1215. [10.1016/j.neuron.2011.02.02](#)
- Dayan, P. (2012). How to set the switches on this thing. *Current Opinion in Neurobiology*, 22 (6), 1068-1074. [10.1016/j.conb.2012.05.011](#)
- Dayan, P. & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, 8 (4), 429-453. [10.3758/CABN.8.4.429](#)
- Edwards, M. J., Adams, R. A., Brown, H., Pareés, I. & Friston, K. (2012). A Bayesian account of 'hysteria'. *Brain*, 135 (11), 3495-3512. [10.1093/brain/aws129](#)
- Egner, T., Monti, J. M. & Summerfield, C. (2010). Expectation and surprise determine neural population responses in the ventral visual stream. *Journal of Neuroscience*, 30 (49), 16601-16608. [10.1523/JNEUROSCI.2770-10.2010](#)
- Estany, A. & Sturm, T. (Eds.) (2014). Extended cognition: New philosophical perspectives. *Special Issue of Philosophical Psychology*, 27 (1)
- Feldman, A. G. (2009). New insights into action-perception coupling. *Experimental Brain Research*, 194 (1), 39-58. [10.1007/s00221-008-1667-3](#)
- Feldman, H. & Friston, K. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, 4 (215), 1-23. [10.3389/fnhum.2010.00215](#)
- Felleman, D. J. & Van Essen, D. C. (1991). Distributed hierarchical processing in primate cerebral cortex. *Cerebral Cortex*, 1 (1), 1-47. [10.1093/cercor/1.1.1-a](#)
- Fink, P. W., Foo, P. S. & Warren, W. H. (2009). Catching fly balls in virtual reality: A critical test of the outfielder problem. *Journal of Vision*, 9 (13), 1-8. [10.1167/9.13.14](#)
- Flash, T. & Hogan, N. (1985). The coordination of arm movements: An experimentally confirmed mathematical model. *The Journal of Neuroscience*, 5 (7), 1688-1703. [10.1.1.134.529](#)
- Fletcher, P. & Frith, C. (2009). Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, 10, 48-58. [10.1038/nrn2536](#)
- Franklin, D. W. & Wolpert, D. M. (2011). Computational mechanisms of sensorimotor control. *Neuron*, 72 (3), 425-442. [10.1016/j.neuron.2011.10.006](#)
- Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences* 29, 360 (1456), 815-836. [10.1098/rstb.2005.1622](#)
- (2008). Hierarchical models in the brain. *PLoS Computational Biology*, 4 (11), e1000211. [10.1371/journal.pcbi.1000211](#)
- (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, 13 (7), 293-301. [10.1016/j.tics.2009.04.005](#)
- (2011). What is optimal about motor control? *Neuron*, 72 (3), 488-498. [10.1016/j.neuron.2011.10.018](#)
- Friston, K. J., Daunizeau, J., Kilner, J. & Kiebel, S. J. (2010). Action and behavior: A free-energy formulation. *Biological Cybernetics*, 102 (3), 227-260. [10.1007/s00422-010-0364-z](#)
- Friston, K., Mattout, J. & Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics*, 104 (1-2), 137-160. [10.1007/s00422-011-0424-z](#)
- Friston, K., Adams, R. A., Perrinet, L. & Breakspear, M. (2012). Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology*, 3 (151), 1-20. [10.3389/fpsyg.2012.00151](#)

- Friston, K., Samothrakis, S. & Montague, R. (2012). Active inference and agency: Optimal control without cost functions. *Biological Cybernetics*, 106 (8-9), 523-541. [10.1007/s00422-012-0512-8](https://doi.org/10.1007/s00422-012-0512-8)
- Friston, K. J., Shiner, T., Fitzgerald, T., Galea, J. M., Adams, R., Brown, H., Dolan, R. J., Moran, R., Stephan, K. E. & Bestmann, S. (2012). Dopamine, affordance and active inference. *PLoS Computational Biology*, 8 (1), e1002327. [10.1371/journal.pcbi.1002327](https://doi.org/10.1371/journal.pcbi.1002327)
- Frith, C. D. & Friston, K. J. (2012). False perceptions and false beliefs: Understanding schizophrenia. *Working Group on Neurosciences and the Human Person: New Perspectives on Human Activities, The Pontifical academy of Sciences, 8-10 November 2012*. Vatican City, VA: Casina Pio IV.
- Gläscher, J., Daw, N., Dayan, P. & O'Doherty, J. P. (2010). States versus rewards: Dissociable neural prediction error signals underlying model based and model-free reinforcement learning. *Neuron*, 66 (4), 585-595. [10.1016/j.neuron.2010.04.016](https://doi.org/10.1016/j.neuron.2010.04.016)
- Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences*, 27 (3), 377-442. [10.1017/S0140525X04000093](https://doi.org/10.1017/S0140525X04000093)
- Haruno, M., Wolpert, D. M. & Kawato, M. (2003). Hierarchical MOSAIC for movement generation. *International congress series*, 1250, 575-590.
- Hinton, G. E., Osindero, S. & Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18 (7), 1527-1554. [10.1162/neco.2006.18.7.1527](https://doi.org/10.1162/neco.2006.18.7.1527)
- Hinton, G. E. & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313 (5786), 504-507. [10.1126/science.1127647](https://doi.org/10.1126/science.1127647)
- Hohwy, J. (2013). *The predictive mind*. New York, NY: Oxford University Press.
- (2014). The neural organ explains the mind. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Hutchins, E. (1995). *Cognition in the wild*. Cambridge, MA: MIT Press.
- (2014). The cultural ecosystem of human cognition. *Philosophical Psychology*, 27 (1), 34-49. [10.1080/09515089.2013.830548](https://doi.org/10.1080/09515089.2013.830548)
- James, W. (1890). *The principles of psychology Vol. I, II*. Cambridge, MA: Harvard University Press.
- Jordan, M. I. & Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6 (2), 181-214. [10.1162/neco.1994.6.2.181](https://doi.org/10.1162/neco.1994.6.2.181)
- Kahneman, D. (2011). *Thinking fast and slow*. London, UK: Penguin.
- Kawato, K. (1999). Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology*, 9 (6), 718-727. [10.1016/S0959-4388\(99\)00028-8](https://doi.org/10.1016/S0959-4388(99)00028-8)
- Land, M. (2001). Does steering a car involve perception of the velocity flow field? In J. M. Zanker & J. Zeil (Eds.) *Motion vision - Computational, neural, and ecological constraints* (pp. 227-238). Berlin, GER: Springer Verlag.
- Landy, D. & Goldstone, R. L. (2005). How we learn about things we don't already understand. *Journal of Experimental and Theoretical Artificial Intelligence*, 17 (4), 343-369. [10.1080/09528130500283832](https://doi.org/10.1080/09528130500283832)
- Littman, M., Majercik, S. & Pitassi, T. (2001). Stochastic Boolean satisfiability. *Journal of Automated Reasoning*, 27 (3), 251-296.
- Lotze, H. (1852). *Medizinische Psychologie oder Physiologie der Seele*. Leipzig, GER: Weidmannsche Buchhandlung.
- Marr, D. (1982). *Vision: A computational approach*. San Francisco, CA: Freeman & Co.
- Maturana, H. & Varela, F. (1980). *Autopoiesis and cognition: The realization of the living*. Boston, MA: Reidel.
- McClelland, J. L. (2013). Integrating probabilistic models of perception and interactive neural networks: A historical and tutorial review. *Frontiers in Psychology*, 4 (503), 1-25. [10.3389/fpsyg.2013.00503](https://doi.org/10.3389/fpsyg.2013.00503)
- Menary, R. (Ed.) (2010). *The extended mind*. Cambridge, MA: MIT Press.
- Mohan, V., Morasso, P., Metta, G. & Kasderidis, S. (2010). Actions & imagined actions in cognitive robots. In V. Cutsuridis, A. Hussain & J. G. Taylor (Eds.) *Perception-reason-action cycle: Models, architectures, and hardware* (pp. 1-32). New York, NY: Springer Series in Cognitive and Neural Systems.
- Mohan, V. & Morasso, P. (2011). Passive motion paradigm: An alternative to optimal control. *Frontiers in Neurorobotics*, 5 (4), 1-28. [10.3389/fnbot.2011.00004](https://doi.org/10.3389/fnbot.2011.00004)
- Namikawa, J., Nishimoto, R. & Tani, J. (2011). A neurodynamic account of spontaneous behaviour. *PLoS Computational Biology*, 7 (10), e100222. [10.1371/journal.pcbi.1002221](https://doi.org/10.1371/journal.pcbi.1002221)
- Namikawa, J. & Tani, J. (2010). Learning to imitate stochastic time series in a compositional way by chaos. *Neural Networks*, 23 (5), 625-638. [10.1016/j.neunet.2009.12.006](https://doi.org/10.1016/j.neunet.2009.12.006)
- Park, J. C., Lim, J. H., Choi, H. & Kim, D. S. (2012). Predictive coding strategies for developmental neurorobotics. *Frontiers in Psychology*, 3 (134), 1-10. [10.3389/fpsyg.2012.00134](https://doi.org/10.3389/fpsyg.2012.00134)

- Pellicano, E. & Burr, D. (2012). When the world becomes too real: A Bayesian explanation of autistic perception. *Trends in Cognitive Sciences*, 16 (10), 504-510. [10.1016/j.tics.2012.08.009](https://doi.org/10.1016/j.tics.2012.08.009)
- Pezzulo, G. (2008). Coordinating with the future: The anticipatory nature of representation. *Minds and Machines*, 18, 179-225. [10.1007/s11023-008-9095-5](https://doi.org/10.1007/s11023-008-9095-5)
- (2014). Why do you fear the bogeyman? An embodied predictive coding model of perceptual inference. *Cognitive, Affective, & Behavioral Neuroscience*, 14 (3), 902-911. [10.3758/s13415-013-0227-x](https://doi.org/10.3758/s13415-013-0227-x)
- Pezzulo, G., Barsalou, L., Cangelosi, A., Fischer, M., McRae, K. & Spivey, M. (2013). Computational grounded cognition: A new alliance between grounded cognition and computational modeling. *Frontiers in Psychology*, 3 (612), 1-11. [10.3389/fpsyg.2012.00612](https://doi.org/10.3389/fpsyg.2012.00612)
- Pezzulo, G., Rigoli, F. & Chersi, F. (2013). The mixed instrumental controller: Using value of information to combine habitual choice and mental simulation. *Frontiers in Psychology*, 4 (92), 1-15. [10.3389/fpsyg.2013.00092](https://doi.org/10.3389/fpsyg.2013.00092)
- Poeppel, D. & Monahan, P. J. (2011). Feedforward and feedback in speech perception: Revisiting analysis by synthesis. *Language and Cognitive Processes*, 26 (7), 935-951. [10.1080/01690965.2010.493301](https://doi.org/10.1080/01690965.2010.493301)
- Price, C. J. & Devlin, J. T. (2011). The interactive Account of ventral occipito-temporal contributions to reading. *Trends in Cognitive Sciences*, 15 (6), 246-253. [10.1016/j.tics.2011.04.001](https://doi.org/10.1016/j.tics.2011.04.001)
- Raichle, M. E. & Snyder, A. Z. (2007). A default mode of brain function: A brief history of an evolving idea. *NeuroImage*, 37 (4), 1083-1090. [10.1016/j.neuroimage.2007.02.041](https://doi.org/10.1016/j.neuroimage.2007.02.041)
- Rao, R. & Ballard, D. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2 (1), 79-87. [10.1038/4580](https://doi.org/10.1038/4580)
- Roepstorff, A. (2013). Interactively human: Sharing time, constructing materiality: Commentary on Clark. *Behavioral and Brain Sciences*, 36 (3), 224-225. [10.1017/S0140525X12002427](https://doi.org/10.1017/S0140525X12002427)
- Roepstorff, A., Niewöhner, J. & Beck, S. (2010). Enculturating brains through patterned practices. *Neural Networks*, 23, 1051-1059. [10.1016/j.neunet.2010.08.002](https://doi.org/10.1016/j.neunet.2010.08.002)
- Roth, M. J., Synofzik, M. & Lindner, A. (2013). The cerebellum optimizes perceptual predictions about external sensory events. *Current Biology*, 23 (10), 930-935. [10.1016/j.cub.2013.04.027](https://doi.org/10.1016/j.cub.2013.04.027)
- Rupert, R. (2004). Challenges to the hypothesis of extended cognition. *Journal of Philosophy*, 101 (8), 389-428.
- (2009). *Cognitive systems and the extended mind*. Oxford, UK: Oxford University Press.
- Saegusa, R., Sakka, S., Metta, G. & Sandini, G. (2008). *Sensory prediction learning - how to model the self and environment*. Annecy, FR: The 12th IMEKO TC1-TC7 joint Symposium on “Man Science and Measurement” (IMEKO2008).
- Seth, A. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, 17 (11), 565-573. [10.1016/j.tics.2013.09.007](https://doi.org/10.1016/j.tics.2013.09.007)
- Shaffer, D. M., Krauchunas, S. M., Eddy, M. & McBeath, M. K. (2004). How dogs navigate to catch frisbees. *Psychological Science*, 15 (7), 437-441. [10.1111/j.0956-7976.2004.00698.x](https://doi.org/10.1111/j.0956-7976.2004.00698.x)
- Shea, N. (2013). Perception vs. action: The computations may be the same but the direction of fit differs: Commentary on Clark. *Behavioral and Brain Sciences*, 36 (3), 228-229. [10.1017/S0140525X12002397](https://doi.org/10.1017/S0140525X12002397)
- Shipp, S., Adams, R. A. & Friston, K. J. (2013). Reflections on agranular architecture: Predictive coding in the motor cortex. *Trends in Neurosciences*, 36 (12), 706-716. [10.1016/j.tins.2013.09.004](https://doi.org/10.1016/j.tins.2013.09.004)
- Smith, L. & Gasser, M. (2005). The development of embodied cognition: Six lessons from babies. *Artificial Life*, 11, 13-29. [10.1162/1064546053278973](https://doi.org/10.1162/1064546053278973)
- Sommer, M. A. & Wurtz, R. H. (2006). Influence of thalamus on spatial visual processing in frontal cortex. *Nature*, 444 (7117), 374-377. [10.1038/nature05279](https://doi.org/10.1038/nature05279)
- (2008). Brain circuits for the internal monitoring of movements. *Annual Review of Neuroscience*, 31 (1), 317-338. [10.1146/annurev.neuro.31.060407.125627](https://doi.org/10.1146/annurev.neuro.31.060407.125627)
- Spivey, M. J. (2007). *The continuity of mind*. New York, NY: Oxford University Press.
- Spratling, M. W. (2008). Predictive coding as a model of biased competition in visual attention. *Annual Review of Neuroscience*, 48 (12), 1391-1408. [10.1146/annurev.neuro.31.060407.125627](https://doi.org/10.1146/annurev.neuro.31.060407.125627)
- (2010). Predictive coding as a model of response properties in cortical area V1. *The Journal of Neuroscience*, 30 (9), 3531-3543. [10.1523/JNEUROSCI.4911-09.2010](https://doi.org/10.1523/JNEUROSCI.4911-09.2010)
- (2013). Distinguishing theory from implementation in predictive coding accounts of brain function. *Behavioral and Brain Sciences*, 36 (3), 231-232.
- Stigler, J. W. (1984). “Mental abacus”: The effect of abacus training on Chinese children mental calculation. *Cognitive Psychology*, 16 (2), 145-176. [10.1016/0010-0285\(84\)90006-9](https://doi.org/10.1016/0010-0285(84)90006-9)
- Tani, J. (2007). On the interactions between top-down anticipation and bottom-up regression. *Frontiers in Neurobotics*, 1 (2), 2. [10.3389/neuro.12.002.2007](https://doi.org/10.3389/neuro.12.002.2007)

- Thelen, E. & Smith, L. (1994). *A dynamic systems approach to the development of cognition and action*. Massachusetts, MA: MIT Press.
- Todorov, E. (2004). Optimality principles in sensorimotor control. *Nature Neuroscience*, 7 (9), 907-915.
[10.1038/nm1309](https://doi.org/10.1038/nm1309)
- Todorov, E. & Jordan, M. I. (2002). Optimal feedback control as a theory of motor coordination. *Nature Neuroscience*, 5 (11), 1226-1235. [10.1038/nm963](https://doi.org/10.1038/nm963)
- Tomasello, M., Carpenter, M., Call, J., Behne, T. & Moll, H. (2005). Understanding and sharing intentions: The ontogeny and phylogeny of cultural cognition. *Behavioral and Brain Sciences*, 28 (5), 675-691.
[10.1017/S0140525X05000129](https://doi.org/10.1017/S0140525X05000129)
- Uno, Y., Kawato, M. & Suzuki, R. (1989). Formation and control of optimal trajectory in human multijoint arm movement. *Biological Cybernetics*, 61 (2), 89-101.
[10.1007/BF00204593](https://doi.org/10.1007/BF00204593)
- Von Holst, E. (1954). "Relations between the central Nervous System and the peripheral organs". *The British Journal of Animal Behaviour*, 2 (3), 89-94.
[10.1016/S0950-5601\(54\)80044-X](https://doi.org/10.1016/S0950-5601(54)80044-X)
- Wolpert, D. M., Doya, K. & Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society of London*, 358 (1431), 593-602.
[10.1098/rstb.2002.1238](https://doi.org/10.1098/rstb.2002.1238)
- Wolpert, D. M. & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks*, 11 (7-8), 1317-1329.
[10.1016/S0893-6080\(98\)00066-5](https://doi.org/10.1016/S0893-6080(98)00066-5)
- Wolpert, M. & Miall, D. M. (1996). Forward Models for Physiological Motor Control. *Neural Networks*, 9 (8), 1265-1279.
- Wurtz, R. H., McAlonan, K., Cavanaugh, J. & Berman, R. A. (2011). Thalamic pathways for active vision. *Trends in Cognitive Sciences*, 15 (4), 177-184.
[10.1016/j.tics.2011.02.004](https://doi.org/10.1016/j.tics.2011.02.004)
- Yamashita, Y. & Tani, J. (2008). Emergence of functional hierarchy in a multiple timescale neural network model: a humanoid robot experiment. *PLoS ONE*, 6 (10), e1000220.
[10.1371/annotation/c580e39c-00bc-43a2-9b15-af71350f9d43](https://doi.org/10.1371/annotation/c580e39c-00bc-43a2-9b15-af71350f9d43)
- Zorzi, M., Testolin, A. & Stoianov, I. (2013). Modeling language and cognition with deep unsupervised learning: a tutorial overview. *Frontiers Psychology*, 4 (415), 1-14. [10.3389/fpsyg.2013.00515](https://doi.org/10.3389/fpsyg.2013.00515)

Extending the Explanandum for Predictive Processing

A Commentary on Andy Clark

Michael Madary

In this commentary, I suggest that the predictive processing framework (PP) might be applicable to areas beyond those identified by Clark. In particular, PP may be relevant for our understanding of perceptual content, consciousness, and for applied cognitive neuroscience. My main claim for each area is as follows:

- 1) PP urges an organism-relative conception of perceptual content.
- 2) Historical *a priori* accounts of the structure of perceptual experience converge with results from PP.
- 3) There are a number of areas in which PP can find important practical applications, including education, public policy, and social interaction.

Keywords

Anticipation | Applied cognitive neuroscience | Consciousness | Perception | Perceptual content | Phenomenology | Predictive processing

Commentator

Michael Madary

madary@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Target Author

Andy Clark

Andy.Clark@ed.ac.uk

University of Edinburgh
Edinburgh, United Kingdom

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

An understandable reaction to the predictive processing framework (PP) is to think that it is too ambitious ([Hohwy this collection](#)). My suggestion in this commentary is the opposite. I will argue that PP can be fruitfully applied to areas of inquiry that have so far received little, if any, attention from the proponents of PP. Perhaps we can extend the explanandum even further than Andy Clark has recommended.

There is a certain rhetorical danger to the position I am urging. One should not oversell

one's case. I hope to avoid this danger by being clear upfront that my goal is not to convince the skeptic of the attraction of PP. I cannot improve on Clark (and others, see below) in that regard. Instead, I investigate the following question: if some version of PP (again, see below) is true, then what are the larger implications for human self-understanding? My answer to this question covers three topics. First I will engage with Clark's discussion of perceptual processing from sections 1 and 2.1 of his article. There I

will sketch how PP's reversal of the traditional model of perceptual processing may have significant implications for the way in which we understand perceptual content, which is a core issue in the philosophy of psychology. In the [second](#) section I will turn to another area of philosophical concern: consciousness. Historically, consciousness research has had a rocky relationship with the sciences of the mind. I hope to point towards the possibility of a rapprochement. In the final section of the commentary, I will quickly touch on some practical matters. If PP is true, then there are important consequences for the way in which we approach topics in education, public policy, and social interaction.

My goal is to indicate possible areas in which Clark's article (and related themes) might serve as a foundation for future directions of research. My main claims are as follows, numbered according to each section:

1. PP urges an organism-relative conception of perceptual content.
2. Historical *a priori* accounts of the structure of perceptual experience converge with results from PP.
3. There are a number of areas in which PP can find important practical applications.

Before entering into the specific issues, I should add a note about what I mean by PP. Here I am following the general theoretical framework expressed in Clark's article as well as in a number of other publications ([Clark 2013](#); [Hohwy 2013](#)). The approach has a number of intellectual roots, including [Hermann von Helmholtz \(1867\)](#) and [Richard Gregory \(1980\)](#). The main contemporary expression of PP perhaps owes the most to [Karl Friston \(2005, 2008, 2010\)](#) and his collaborators, also with important developments of the generative model by [Geoffrey Hinton \(2007\)](#). By referring to PP as one general framework, I do not mean to imply that there are no outstanding issues of disagreement or open questions within PP. As Clark indicates, citing [Spratling \(2013\)](#), there are a number of options being developed as to the specific implementation of PP. Also, in the philosophical lit-

erature there is an emerging question about whether to understand PP as internalist or externalist regarding the vehicles of mental states ([Hohwy 2014](#))—I take no position either way here, but see footnote 2. Overall, my remarks are motivated by Clark's exposition of PP, but they should be applicable to other approaches and interpretations as well.

2 A new conception of perceptual content

Clark has emphasized the way in which PP departs from the standard picture in perceptual psychology, and from [David Marr's \(1982\)](#) model of visual processing in particular (pp. 1–5). According to the standard account, the flow of information is “bottom-up,” as perceptual systems construct increasingly sophisticated representations based on the information transduced at the periphery. According to PP, perception involves the active prediction of the upcoming sensory input, “top-down.” Deviation from what is predicted, known as the prediction error, propagates upwards through the hierarchy until it is explained away by the Bayesian generative model.

Now I would like to add that the standard picture in perceptual psychology has been widely regarded as complementary to the standard picture in the philosophy of perception (see [Tye 2000](#), for example). One central question in the philosophy of perception is the following: what is the *content* of perceptual states? Or, what does perception *represent*? The standard answer, in tune with Marr's approach, is that perceptual systems represent the external world, more or less as it really is. As [Marr](#) puts it, the purpose of vision is “to know what is where by looking” ([1982](#)). This way of thinking about perceptual content is almost a commonplace in the philosophical literature ([Lewis 1980](#), p. 239; [Fodor 1987](#), Ch. 4; [Dretske 1995](#), Ch. 1). [Kathleen Akins](#) has described how the orthodox conception regards the senses as “servile” in that they report on the environmental stimulus “without fiction or embellishment” ([1996](#), pp. 350–351).

Since PP overturns the reigning model in perceptual psychology, one might now ask

whether it also overturns the reigning model in the philosophy of perception. Here are two initial reasons to think that it does. First, according to PP, there is always an active contribution from the organism, or at least from a part of the organism. Perceptual states are generated internally and spontaneously by the ongoing dynamics of the generative model. Those states are *constrained* by perceptual sampling of the world, not driven by input from the world. Perceptual states are driven by the endogenous activity of the predictive brain. The relevant causal history of these states begins, if you will, within the brain, rather than from the outside. Each organism's generative model is unique in that it has been formed and continuously revised according to the particular trajectory of that organism's cycle of action and perception. As Clark himself puts it, the forward flow of sensory information is always "*relative to specific predictions*" (p. 6). These considerations make it clear that there can be variation in perceptual content for identical environmental conditions. Perceivers with different histories will have different predictions (Madary 2013, pp. 342–345). The degree of variation is an open question, but it is reasonable to expect variation.

A second reason to think that PP motivates a richer conception of perceptual content is that perception, according to PP, is not simply in the service of informing the organism "what is where." One main feature of PP is that perception and action work together in the service of minimizing prediction error. Clark explains that in "active inference [...]" the agent moves its sensors in ways that amount to actively seeking or generating the sensory consequences that they [...] expect" (2013, p. 6, also see his discussion on page 16). If this is right, then perception does not serve the purpose of simply reporting on the state of the environment. Instead, perception is guided by expectation. While the received view of perceptual content answers the question of "what is out there?", PP suggests that perceptual content answers the question of "is this what I expected and tested via active inference?" In a way, PP simplifies perceptual content by replacing the goal

of representing the world with the single guiding principle of error minimization.

These two points suggest an understanding of perceptual content as something that is deeply informed by the specific history and embodiment of the organism. The content of perception is a complex interplay between particular organisms and their particular environments. At least on the face of it, this way of considering perception suggests new challenges and interesting new theoretical options for philosophers interested in describing perceptual content. For one thing, it suggests that propositional content as expressed using natural language (Searle 1983, p. 40) may be ill-suited for the task of describing perceptual content. Natural language does not typically include reports about prediction-error minimization, nor does it capture the fine-grained differences in perceptual content that will arise due to slight variations in the predictions made by different organisms. The traditional account of perceptual content, following Marr, does not include such differences, and is thus better disposed to expression using natural language.

These new challenges for understanding perceptual content may offer at the same time a general lesson for understanding all mental content in a naturalistic manner. Let me explain. One of the main goals in the philosophy of psychology has been to naturalize intentionality, to give an account of the content of mental states in terms of the natural sciences (in non-mentalistic terms). Well-known attempts include causal co-variation (Fodor 1987, Ch. 4) and teleosemantics (Millikan 1984, 2004). All attempts have met with compelling counterexamples.¹ Importantly, one implicit presupposition in the debate is that mental content should be conceived along the lines of the traditional view of perceptual content sketched above. That is, mental states are thought to be about bits of the objective world considered independently of the particular organism who possesses those mental states. To use a standard example, my belief that there is milk in the refrigerator is true if and only if there is milk in the refriger-

¹ For an overview of the major theories and their challenges, see Jacob (2010, section 9) and the references therein.

ator. This belief is about bits of the objective world: milk and the refrigerator in particular. Nothing else about my mind is deemed relevant for understanding the content of that belief. To use the familiar phrase, beliefs have a mind-to-world direction of fit (based on [Anscombe 1957](#), §32).

If my reading of PP is right, and perceptual content turns out to be a matter of the complex interaction between particular organisms and their environments, then the comfortable pre-theoretical mind/world distinction might need revision.² Recall the discussion above, in which I claimed that, on the new PP-inspired understanding of perception the question is about whether sensory stimulation fulfils the expectations of particular organisms. All perceptual states are thereby colored, as it were, by the mental lives of the organisms having those states. Organisms are not interested in what the world is like. Organisms are interested in sustaining their integrity and physical existence; they are interested in what the world is like *relative to their own particular sensorimotor trajectory through the world*, a trajectory that is partly determined by their phenotype ([Friston et al. 2006](#)). This refashioning of the mind/world relationship is unorthodox, but it is hardly new. Similar ideas can be found in [von Uexküll's Umwelt \(1934\)](#), [Merleau-Ponty's](#) discussion of sensory stimuli (1962, p. 79), [Milikan's](#) “pushmi-pullyu” representations (1995), [Akins' narcissistic sensory systems \(1996\)](#), [Clark's](#) earlier work (1997, Ch. 1), and in [Metzinger's](#) ego tunnel (2009, pp. 8–9).

Now return to the problem of naturalizing intentionality. If we replace the notion of a purely world-directed mental state with a world-relative-to-the-organism-directed mental state, then naturalizing intentionality must somehow incorporate the relationship between

the organism and its world. One way to pursue this project is to make it a matter of biology and physics. All living organisms keep themselves far from thermodynamic equilibrium by continuously exchanging matter and energy with their environment ([Haynie 2008](#)). Perhaps intentionality can be recast in terms of the organism's ongoing struggle to maintain itself as a living entity. This line of thought is central to the enactivist “sense-making” of [Maturana, Varela, and Thompson \(Maturana & Varela 1980; Thompson 2007\)](#). Crucially, it is also a central feature of [Friston's](#) version of PP. According to [Friston](#), prediction error minimization is a kind of functional description for the physical process of the organism's minimizing free energy in its effort to maintain itself far from thermodynamic equilibrium (2013). Naturalizing intentionality may be just a matter of physics (see [Dixon et al. 2014](#) for an implementation of this strategy for problem-solving tasks).

Before moving on to the next section, I should add two qualifications. First, the idea of perceptual content being partly determined by the particular history of the perceiver should not be misunderstood as some kind of radical relativism with regard to perceptual content. Even if perceptual content is *partly* determined by the details of the organism, it is also partly determined by the world itself. As proponents of PP frequently claim, our generative models mirror the causal structure of the world ([Hohwy 2013](#), Ch. 1). The point I am emphasizing here is that the causal structure of the world that is extracted is a structure relative to the embodiment (see [Clark this collection](#), section 2.4)—and perceptual history—of the perceiver. The causal structure mirrored by a chimpanzee's generative model is, in important ways, unlike the causal structure mirrored by that of a catfish.

The second qualification has to do with my remark that naturalizing intentionality may be just a matter of physics. Even if one allows that the approach I sketched shows promise, it is important to emphasize the explanatory gulf that remains. The intentionality-as-physics approach might succeed in explaining a bacterium's intentional directedness towards a

² One possibility here has been explored recently by Karl Friston using the concept of a Markov blanket, which produces a kind of partition between information states. As I read Friston, he advocates a pluralism about Markov blankets. On this view, there is not one boundary between mind and world, but instead there are a number of salient boundaries within, and perhaps around, living organisms. [Friston](#) writes that “... a system can have a multitude of partitions and Markov blankets ... the Markov blanket of an animal encloses the Markov blankets of its organs, which enclose Markov blankets of cells, which enclose Markov blankets of nuclei ...” (2013, p. 10).

sugar gradient (Thompson 2007, p. 74–75), but it is far from clear how it would apply to my belief that P—say, for example, that California Chrome won the Kentucky Derby in 2014.

The main argument of this section has been that PP motivates an understanding of perceptual content that is always organism-relative. Clark’s version of PP, while not in conflict with this idea, has not addressed it explicitly, especially as it relates to the philosophy of perception. My goal here has been to do just that.

3 Consciousness

In this section I would like to consider how conscious experience might relate to the PP framework. In particular, I suggest that there is a convergence between *a priori* descriptions of consciousness, on one hand, and the structure of information processing according to PP on the other.³ I will not remark on the way in which PP relates to some well-known issues in the study of consciousness, such as the hard problem or the explanatory gap. It is not clear to me that PP has anything new to contribute to these topics. Nor will I make any claims about which existing theories of the neural basis of consciousness fit best with PP, although I suspect there is some interesting work there to be done.

My main concern here is in the *structure* of conscious experience, of visual experience in particular. Here I adopt a strategy recommended by Thomas Nagel (1974), and David Chalmers (1996, pp. 224–225). Nagel puts the idea nicely, “[...] structural features of perception might be more accessible to objective description, even though something would be left out” (1974, p. 449, cited in Chalmers 1996, pp. 382 f.). The strategy has been implemented, in fact, using Marr’s theory of vision—the theory that, as Clark puts it, PP turns upside down. Ray Jackendoff (1987, p. 178) and Jesse Prinz (2012, p. 52) have both emphasized the structural similarities between conscious visual experience and Marr’s 2.5 dimensional sketch.

³ For a theoretical treatment of the functional significance of this convergence, see Metzinger & Gallese (2003).

Visual phenomenology is not a flat two-dimensional surface, because we see depth. But neither is visual phenomenology fully three-dimensional, because we cannot see the hidden sides of objects. Marr’s 2.5 dimensional representation captures the level in-between two and three dimensional representation that seems to correspond to our visual phenomenology; it captures Hume’s insight that visual experience is perspectival: “The table, which we see, seems to diminish, as we remove farther from it [...]” (1993, p. 104).

As Hume emphasized the perspectival nature of visual experience, Kant famously emphasized the temporal nature of experience in the second section of the *Transcendental Aesthetic*: “Time is a necessary representation (*Vorstellung*), which lays at the foundation of all intuitions” (1781/1887/1998, A31). In an elegant synthesis of these two features of visual experience, Edmund Husserl suggested that the general structure of visual experience is one of anticipation and fulfillment:

Every percept, and every perceptual context, reveals itself, on closer analysis, as made up of components which are to be understood as ranged under two stand-points of intention and (actual or possible) fulfillment. (*Logical Investigation*, VI §10 1900, Findlay trans., 1970)

In this passage from his early work, Husserl writes of “intention and fulfillment,” but he later replaced “intention” with “anticipation” when dealing with perception.⁴

The main point is fairly straightforward: we perceive properties by implicitly anticipating how the appearances of those properties will

⁴ When first developing the framework, he used the more general term “intention” because he was dealing with linguistic meaning, not perception. When applying the framework to perception one can be more precise about the nature of the empty perceptual intentions: they are anticipatory. In his later work, his *Analyses of Passive Synthesis* from the 1920s, Husserl ties in perceptual intentions with his work on time consciousness (1969) and refers to them as protentions (*Protentionen*; Husserl 1966, p. 7). In the same work, he refers to perceptual protentions as anticipations (*Erwartungen*, 1966, p. 13, and *antizipiert*, 1966, p. 7). See Madary (2012a) for a discussion of how Husserl’s framework can be situated relative to contemporary philosophy of perception. Also see Bernet et al. (1993, p. 128) and Hopp (2011).

change as we move (or as the objects move). Husserl's proposal accommodates the perspectival character of experience because it addresses the question of how we perceive objective properties despite being constrained to one perspective at a time. And it accommodates the temporal nature of experience because anticipation is always future-directed.

Here is not the place to enter into the details of the thesis that the general structure of conscious experience is one of anticipation and fulfillment (see my 2013 for some of these details), but I should add one more point. As both Husserl (1973, p. 294) and Daniel Dennett (1991, Ch. 3) have noted, peripheral vision is highly indeterminate.⁵ Also, as we explore our environment we experience a continuous trade off between determinacy and indeterminacy. As I lean in for a closer look at one object, the other objects in my visual field fade into indeterminacy. In order to account for this feature of experience, we can note that visual anticipations have various degrees of determinacy.⁶

Now let us return to PP. *If Hume provides the philosophy of perception for Marr's theory of vision, then Husserl provides the philosophy of perception for PP.* The structural similarities should be apparent. The predictive brain underlies the essentially anticipatory structure of perceptual awareness. Degrees of determinacy are encoded probabilistically in our generative models (Clark 2013; Madary 2012b). Action and perception are tightly linked (Clark this collection, p. 9) as self-generated movements stir up new perceptual anticipations.

Many readers will see a connection between the thesis of anticipation and fulfillment, on one hand, and the sensorimotor approach to perception (O'Regan & Noë 2001; Noë 2004) on the other. Overall, there is significant thematic overlap between the two (Madary 2012a, p. 149). As Seth (2014) has argued, many of the central claims of the sensorimotor approach can be incorporated into the PP framework.⁷ This synthesis offers impressive explanatory power, bringing the standard sensorimotor experimental evidence (reversing

goggles, change blindness, selective rearing) together with the theoretical neuroscience of PP. The explanatory power is even more impressive if I am correct that PP reflects the general structure of visual phenomenology, where predictive processing corresponds to perceptual anticipations and probabilistic coding corresponds to experienced indeterminacy.

4 Applied cognitive neuroscience

I would like to begin this section with some general comments about new opportunities for human self-understanding, about extending the explanandum. Academic disciplines are standardly divided into the sciences and the humanities, and some have expressed discomfort about the distance between the two modes of inquiry, or between the two cultures, as Snow (1959) famously put it (also see Brockman 1996). There is an immediate appeal to Metzinger's assertion that "Epistemic progress in the real world is something that is achieved by all disciplines together" (2003, p. 4). *If my claims from the previous section are on the right track, then we have a convergence of results between the two independent modes of inquiry, between the empirical sciences and the humanities.* It is tempting to hope that this convergence signals the beginning of a rapprochement between the sciences and the humanities. Perhaps we are at the threshold of a new science of the mind (Rowlands 2010), a science that finds natural and fruitful connections with the world of human experience. In this section, I will explore possible connections with education, public policy, and social interaction.

Clark makes two main claims in the final sections of his article that serve for the basis of my comments here. First, he suggests that PP motivates an understanding of cognitive processing as "maximally context sensitive" (p. 16), which follows from the property of PP systems being highly flexible in setting precision weightings for the incoming prediction errors. Flexibility in weighting precision enables flexibility in the deployment of processing resources. Thus there may be a wide variety of cognitive strategies at our disposal, with a continuous in-

⁵ For impressive empirical work on this theme, see Freeman & Simoncelli (2011).

terplay between more costly and less costly strategies. Second, he addresses the challenge of explaining why humans have unique cognitive powers unavailable to non-human animals who have the same fundamental PP architecture. In response to this challenge, Clark suggests that our abilities may be due to our patterns of social interaction as well as our construction of artifacts and “designer environments” (p. 19). Taken together, these two claims can be used to inform practical decisions in a number of ways.

Begin with education. Educational psychology is a broad and important area of research. PP suggests new ways of approaching human learning, ways that might depart from the received views that have guided educational psychology. I cannot begin to engage with this huge issue here, but I would like to offer one quick example. One fairly well-known application of educational psychology is in the concept of scaffolded learning, which is built on work by Lev Vygotsky and Jerome Bruner. As it is used now, scaffolded learning involves providing the student with helpful aids at particular stages of the learning process. These aids could include having a teacher present to give helpful hints, working in small groups, and various artifacts designed with the intention of anticipating stages at which the student will need help, such as visual aids, models, or tools. Clark himself mentions the abacus, which is central example of scaffolded learning (p. 19). More generally, scaffolded learning is a good example of what Richard Menary has called “cognitive practices,” which he defines as “manipulations of an external representation to complete a cognitive task” (2010, p. 238).

If PP is right, then the learning process could be optimized by designing environments in order to provide the cycle of action and perception with precisely controlled feedback (prediction error). With the growing commercial availability of immersive virtual reality equipment, educators could design learning environments (or help students design their own environments) without the messy constraints of the physical world. PP may give us a framework with which to understand—and predict—the detailed bodily movements of subjects as they

attempt to minimize their own prediction error. Using this framework, we can design systems that would optimize skill acquisition by efficiently predicting the errors that learners will make. This method could be fruitfully applied in the abstract (mathematics), the concrete (skiing), and in-between (foreign languages). Along these lines, the insights of PP, together with emerging technology, can lead to powerful new educational techniques.

Psychology is also applied in some areas of public policy. Clark mentions that PP challenges Kahneman’s well-known model of human thinking as consisting of a fast automatic system and a slower deliberative system (p. 18). Kahneman’s model has been applied as a basis for influential recommendations about laws and public policy in the United States (Thaler & Sunstein 2008; Sunstein 2014). If PP homes in on a more accurate model of the thinking process, then we ought to use it, rather than (or as a complement to?) the dual systems model as a basis for policy making. Clark’s interpretation of PP suggests that we have a highly flexible range of cognitive systems, not limited to Kahneman’s two.

For example, one application of Kahneman’s model might involve the installation of environmental elements meant to appeal to the fast thinking system, to “nudge” agents towards making decisions in their best interest. If Clark is correct, we might consider even more sophisticated environmental features that have the goal of helping agents to deploy their range of cognitive strategies more efficiently. Clark’s ideas of context sensitivity and designer environments are both relevant here. As a society we may wish somehow to create environments and contexts that take advantage of the large repertoire of cognitive strategies available to us, according to Clark’s version of PP (see Levy 2012, for example).

The final topic I’d like to mention in this section is what is best described in general terms as social interaction. I mean to indicate a number of related topics here, but the main issue is how PP might relate to the well-known philosophical topic of the way in which we understand and explain our behavior to one an-

other. Recall, for instance, [Donald Davidson's](#) (1963) claim that our explanation of our behavior in terms of reasons is a kind of causal explanation—reasons as causes. On his influential view, the connection between reason and actions is a causal connection. In contrast, recall [Paul Churchland's](#) envisioning of the golden age of psychology in which we dispose of folk psychological reason-giving in favor of more precise neurophysiological explanations of behavior (1981). According to Churchland's radical alternative, the causes of actions are not reasons as expressed using natural language. Instead, our actions are caused by patterns of neurons firing, patterns that can be described using mathematical tools such as a multidimensional state space. In opposition to Churchland's grand vision, we have [Jerry Fodor's](#) claim that the realization of such a vision would be “the greatest intellectual catastrophe in the history of our species” (1987, p. xii). Is PP the beginning of Churchland's grand vision coming to pass? Is a great intellectual catastrophe looming?

On one hand, PP seems like an obvious departure from folk psychology: Try explaining your X-ing to someone by claiming that you X-ed in order to minimize prediction error! One big issue here will be the way in which we think about agency itself. It seems mistaken to say that minimizing prediction error is something done by an agent. Such a process seems to be better described as occurring sub-personally. On the other hand, it is not inconceivable that propositional attitudes can capture the dynamics of prediction error minimization on a suitably coarse-grained level, perhaps along the lines suggested using symbolic dynamics ([Dale & Spivey 2005](#); [Atmanspacher & beim Graben 2007](#); [Spivey 2007](#), Ch. 10). I suggest that these fascinating issues warrant further investigation. In particular, further investigation ought to incorporate Clark's ideas of maximal context sensitivity and the importance of designer environments.

The way in which we understand each other's behavior is also directly relevant for moral responsibility. Following Peter Strawson's seminal “[Freedom and Resentment](#)” (1962),

philosophers have started thinking about moral responsibility in terms of our reactions to one another, reactions that involve holding each other accountable. On one influential view, we hold each other accountable when our actions issue from our own reasons-responsive mechanisms ([Fischer & Ravizza 1998](#)). On a more recent proposal, holding each other accountable is best modeled as a kind of conversation ([McKenna 2012](#)). These proposals depend, in important ways, on assumptions about human psychology. In particular, they depend on our practice of giving reasons for behavior. As PP suggests a new fundamental underlying principle of behavior, our practices of holding each other accountable may be approached from a new perspective. The new challenge in this area will be to reconcile (if possible) the practice of giving reasons, on one hand, with PP's account of behavior in terms of error minimization on the other.

5 Conclusion

The main theme of my commentary might appear to be driven by an overexcited optimism for the new theory. To be clear, I have not claimed that PP is correct. Even its main proponents are quick to point out that important open issues remain. My claim is that it is worthwhile to consider the full implications of PP, given the convincing evidence presented so far. In this commentary, I have tried to suggest some of the implications that have not yet been mentioned—implications for perceptual content, consciousness, and applied cognitive neuroscience. These implications can be summarized as follows:

1. PP urges an organism-relative conception of perceptual content.
2. Historical *a priori* accounts of the structure of perceptual experience converge with results from PP.
3. There are a number of areas in which PP can find important practical applications.

The final section includes some challenges for future research. The main challenge is one that

has been familiar in one form or another for several decades in the philosophy of mind. This challenge is to address the tension between the way in which we understand and explain our behavior using natural language, on one hand, and our best theory of human behavior from cognitive neuroscience, which, arguably, is PP, on the other hand. In closing I should note that even if key elements of PP are eventually rejected, it might still turn out that our best model of the mind supports some of the themes I have been discussing.

Acknowledgments

I thank Thomas Metzinger and Jennifer Windt for helpful detailed comments on an earlier draft. This research was supported by the EC Project VERE, funded under the EU 7th Framework Program, Future and Emerging Technologies (Grant 257695).

References

- Akins, K. (1996). Of sensory systems and the “aboutness” of mental states. *Journal of Philosophy*, 93 (7), 337-372. [10.2307/2941125](https://doi.org/10.2307/2941125)
- Anscombe, G. E. M. (1957). *Intention*. Oxford, UK: Basil Blackwell.
- Atmanspacher, H. & beim Graben, P. (2007). Contextual emergence of mental states from neurodynamics. *Chaos and Complexity Letters*, 2 (2-3), 151-168.
- Bernet, R., Kern, I. & Marbach, E. (1993). *An introduction to Husserlian phenomenology*. Evanston, IL: Northwestern University Press.
- Brockman, J. (1996). *Third culture: Beyond the scientific revolution*. New York, NY: Touchstone.
- Chalmers, D. (1996). *The conscious mind*. Oxford, UK: Oxford University Press.
- Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy*, 78 (2), 67-90. [10.1111/j.1467-9973.1992.tb00550.x](https://doi.org/10.1111/j.1467-9973.1992.tb00550.x)
- Clark, A. (1997). *Being there: Putting brain, body, and world together again*. Cambridge, MA: MIT Press.
- (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Science*, 36 (3), 1-73. [10.1017/S0140525X12000477](https://doi.org/10.1017/S0140525X12000477)
- (2015). Embodied Prediction. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a.M., GER: MIND Group.
- Dale, R. & Spivey, M. (2005). From apples and oranges to symbolic dynamics: A framework for conciliating notions of cognitive representation. *Journal of Experimental and Theoretical Artificial Intelligence*, 17 (4), 317-342. [10.1080/09528130500283766](https://doi.org/10.1080/09528130500283766)
- Davidson, D. (1963). Actions, reasons and causes. *Journal of Philosophy*, 60 (23), 685-700. [10.2307/2023177](https://doi.org/10.2307/2023177)
- Dennett, D. (1991). *Consciousness explained*. Boston, MA: Little, Brown, & Co.
- Dixon, J., Kelty-Stephen, D. & Anastas, J. (2014). The embodied dynamics of problem solving: New structure from multiscale interactions. In L. Shapiro (Ed.) *The Routledge handbook of embodied cognition*. London, UK: Routledge.
- Dretske, F. (1995). *Naturalizing the mind*. Cambridge, MA: MIT Press.
- Fischer, J. M. & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge, UK: Cambridge University Press.

- Fodor, J. (1987). *Psychosemantics*. Cambridge, MA: MIT Press.
- Freeman, J. & Simoncelli, E. (2011). *Metamers of the ventral stream*. . [10.1038/nrn.2889](https://doi.org/10.1038/nrn.2889)
- Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B*, 360 (1456), 815-836. [10.1098/rstb.2005.1622](https://doi.org/10.1098/rstb.2005.1622)
- (2008). Hierarchical models in the brain. *PLoS Comput Biol*, 4 (11), e1000211. [10.1371/journal.pcbi.1000211](https://doi.org/10.1371/journal.pcbi.1000211)
- (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11 (2), 127-138. [10.1038/nrn2787](https://doi.org/10.1038/nrn2787)
- (2013). Life as we know it. *Journal of the Royal Society Interface*, 10 (86), 1-12. [10.1098/rsif.2013.0475](https://doi.org/10.1098/rsif.2013.0475)
- Friston, K., Kilner, J. & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology - Paris*, 100, 70-87. [10.1016/j.jphysparis.2006.10.001](https://doi.org/10.1016/j.jphysparis.2006.10.001)
- Gregory, R. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society B*, 290 (1038), 181-197.
- Haynie, D. (2008). *Biological thermodynamics*. Cambridge, UK: Cambridge University Press.
- Hinton, G. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences*, 11 (10), 428-34. [10.1016/j.tics.2007.09.004](https://doi.org/10.1016/j.tics.2007.09.004)
- Hohwy, J. (2013). *The predictive mind*. Oxford, UK: Oxford University Press.
- (2014). The self-evidencing brain. *Noûs*, 1-27. [10.1111/nous.12062](https://doi.org/10.1111/nous.12062)
- Hopp, W. (2011). *Perception and knowledge: A phenomenological account*. Cambridge, UK: Cambridge University Press.
- Hume, D. (1993). *An enquiry concerning human understanding*. Indianapolis, IN: Hackett Publishing.
- Husserl, E. (1900). *Logische Untersuchungen*. London, UK: Routledge.
- (1966). *Husserliana XI Analysen zur passiven Synthesis*. Dordrecht, NL: Kluwer.
- (1969). *Husserliana X zur Phänomenologie des inneren Zeitbewusstseins (1893-1917)*. Dordrecht, NL: Kluwer.
- (1973). *Husserliana XVI Ding und Raum: Vorlesungen 1907*. Dordrecht, NL: Kluwer.
- Jackendoff, R. (1987). *Consciousness and the computational mind*. Cambridge, MA: MIT Press.
- Jacob, P. (2010). Intentionality. *Stanford Encyclopedia of Philosophy*, Fall. <http://plato.stanford.edu/entries/intentionality/>
- Kant, I. (1998). *Kritik der reinen Vernunft*. Hamburg, GER: Meiner Verlag.
- Levy, N. (2012). Ecological engineering: Reshaping our environments to achieve our goals. *Philosophy and Technology*, 25 (4), 589-604.
- Lewis, D. (1980). Veridical hallucination and prosthetic vision. *Australasian Journal of Philosophy*, 58 (3), 239-249. [10.1080/00048408012341251](https://doi.org/10.1080/00048408012341251)
- Madary, M. (2012a). Husserl on Perceptual Constancy. *European Journal of Philosophy*, 20 (1), 145-165. [10.1111/j.1468-0378.2010.00405.x](https://doi.org/10.1111/j.1468-0378.2010.00405.x)
- (2012b). How would the world look if it looked as if it were encoded as an intertwined set of probability density distributions? *Frontiers in Psychology*, 3. [10.3389/fpsyg.2012.00419](https://doi.org/10.3389/fpsyg.2012.00419)
- (2013). Anticipation and variation in visual content. *Philosophical Studies*, 165, 335-347. [10.1007/s11098-012-9926-3](https://doi.org/10.1007/s11098-012-9926-3)
- Marr, D. (1982). *Vision: A computational approach*. New York, NY: Freeman and Co.
- Maturana, H. R. & Varela, F. J. (1980). *Autopoiesis and cognition: The realization of the living*. Dordrecht, NL: Reidel.
- McKenna, M. (2012). *Conversation and responsibility*. Oxford, UK: Oxford University Press.
- Menary, R. (2010). Cognitive integration and the extended mind. In R. Menary (Ed.) *The extended mind*. Cambridge, MA: MIT Press.
- Merleau-Ponty, M. (1962). *Phenomenology of perception*. London, UK: Routledge.
- Metzinger, T. (2003). *Being no one: The self-model theory of subjectivity*. Cambridge, MA: MIT Press.
- (2009). *The ego tunnel*. New York, NY: Basic Books.
- Metzinger, T. & Gallese, V. (2003). The emergence of a shared action ontology: Building blocks for a theory. *Consciousness and Cognition*, 12 (4), 549-571. [10.1016/S1053-8100\(03\)00072-2](https://doi.org/10.1016/S1053-8100(03)00072-2)
- Millikan, R. (1984). *Language, thought and other biological objects*. Cambridge, MA: MIT Press.
- (1995). Pushmi-pullyu representations. In J. Tomberlin (Ed.) *Philosophical perspectives 9: AI, connectionism, and philosophical psychology*. Atascadero, CA: Ridgeview Publishing Company.
- (2004). *Varieties of meaning: The 2002 Jean-Nicod Lectures*. Cambridge, MA: MIT Press.
- Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, 83 (4), 435-450.
- Noë, A. (2004). *Action in perception*. Cambridge, MA: MIT Press.

- O'Regan, K. & Noë, A. (2001). A sensorimotor approach to vision and visual consciousness. *Behavioral and Brain Sciences*, 24, 939-973.
- Prinz, J. (2012). *The conscious brain: How attention engenders experience*. Oxford, UK: Oxford University Press.
- Rowlands, M. (2010). *The new science of the mind*. Cambridge, MA: MIT Press.
- Searle, J. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge, MA: Cambridge University Press.
- Seth, A. (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive Neuroscience*, 5 (2), 97-118.
[10.1080/17588928.2013.877880](https://doi.org/10.1080/17588928.2013.877880)
- Snow, C. P. (1959). *The two cultures*. Cambridge, UK: Cambridge University Press.
- Spivey, M. (2007). *The continuity of mind*. Oxford, UK: Oxford University Press.
- Spratling, M. (2013). Distinguishing theory from implementation in predictive coding accounts of brain function. *Behavioral and Brain Sciences*, 36 (3), 51-52.
[10.1017/S0140525X12002178](https://doi.org/10.1017/S0140525X12002178)
- Strawson, P. F. (1962). Freedom and Resentment. *Proceedings of the British Academy*, 48, 1-25.
- Sunstein, C. (2014). *Why nudge?: The politics of libertarian paternalism*. New Haven, CT: Yale University Press.
- Thaler, R. & Sunstein, C. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. New Haven, CT: Yale University Press.
- Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind*. Cambridge, MA: Harvard University Press.
- Tye, M. (2000). *Consciousness, color, and content*. Cambridge, MA: MIT Press.
- von Helmholtz, H. (1867). *Handbuch der Physiologischen Optik*. Leipzig, GER: Leopold Voss.
- von Uexküll, J. (1934). A stroll through the worlds of animals and men. In K. Lashley (Ed.) *Instinctive behavior*. New York, NY: International Universities Press.

Predicting Peace: The End of the Representation Wars

A Reply to Michael Madary

Andy Clark

Michael Madary's visionary and incisive commentary brings into clear and productive focus some of the deepest, potentially most transformative, implications of the Predictive Processing (PP) framework. A key thread running through the commentary concerns the active and "organism-relative" nature of the inner states underlying perception and action. In this Reply, I pick up this thread, expanding upon some additional features that extend and underline Madary's point. I then ask, What remains of the bedrock picture of inner states bearing familiar representational contents? The answer is not clear-cut. I end by suggesting that we have here moved so far from a once-standard complex of ideas concerning the nature and role of the inner states underlying perception and action that stale old debates concerning the existence, nature, and role of "internal representations" should now be abandoned and peace declared.

Keywords

Action | Action-oriented perception | Content | Enaction | Intentionality | Perception | Perceptual content | Predictive coding | Predictive processing | Representation

1 Organism-relative content

I'm hugely indebted to Michael Madary for his visionary and incisive commentary. The commentary covers three topics – the nature of perceptual content, the structure of experience, and some practical implications of the PP (Predictive Processing) framework. Each one deserves a full-length paper in reply, but I will restrict these brief comments to the first topic – the nature of perceptual content. Should the PP vision prove correct, Madary suggests, this would transform our understanding of the nature and role of perceptual content, with potential consequences for the larger project of naturalizing mental content.

Driving such sweeping and radical reform is (Madary argues) the PP emphasis upon the active contribution of the organism to the generation of perceptual states. There is an active contribution, Madary ([this collection](#), section 2) suggests, insofar as PP depicts perceptual states as "generated internally and spontaneously by the internal dynamics of the generative model" (p. 3).

Such a claim clearly requires careful handling. For even the most staunchly feedforward model of perception requires a substantial contribution from the organism. It is thus the nature, not the existence, of that contribution

Author

[Andy Clark](#)

Andy.Clark@ed.ac.uk

University of Edinburgh
Edinburgh, United Kingdom

Commentator

[Michael Madary](#)

madary@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

that must make the difference. Elaborating upon this, Madary notes that ongoing endogenous activity plays a leading role in the PP story. One might say: the organism's generative model (more on which later) is already active, attempting to predict the incoming sensory flow. The flow of incoming information is thus rapidly flipped into a flow tracking "unexpected salient deviation". Identical inputs may thus result in very different perceptual states as predictions alter and evolve. An important consequence, highlighted by Madary, is that different histories of interaction will thus result in different perceptual contents being computed for the very same inputs. Different species, different niches, differences of bodily form, and differences of proximal goals and of personal history are all thus apt (to varying degrees) to transform what is being predicted, and hence the contents properly delivered by the perceptual process.

Those contents are further transformed by a second feature of the PP account: the active selection of perceptual inputs. For at the most fundamental level, the PP story does not depict perception as a process of building a representation of the external world at all. Instead, it depicts perception as just one part of a cohesive strategy for keeping an organism within a kind of "window of viability". To this end the active organism both predicts *and selects* the evolving sensory flow, moving its body and sensory organs so as to expose itself to the sensory stimulations that it predicts. In this way, some of our predictions act as self-fulfilling prophecies, enabling us to harvest the predicted sensory streams. These two features (endogenous activity and the self-selection of the sensory flow) place PP just about maximally distant from traditional, passive "feedforward hierarchy" stories. They are rather (as Mike Anderson once commented to me) the ultimate expression of the "active perception" program.

Here too, though, we should be careful to nuance our story correctly. For part of maintaining ourselves in a long-term window of viability may involve not just seeking out the sensory flows we predict, but the active elicitation of many that we don't! PP may, in fact, man-

date all manner of short-term explorations and self-destabilizations. But such delicacies (though critically important- see [Clark \(in press\)](#) chapters 8 and 9) may safely be left for another day. The present upshot ([Madary this collection](#), section 2) is simply that PP, instead of depicting perception as a mechanism for revealing "what is where" in the external world, turns out to be a mechanism for engaging the external world in ways that say as much about the organism (and its own history) as they do about the world outside. To naturalize intentionality, then, "all" we need do is display the mechanisms by which such ongoing viability-preserving engagements are enabled, and make intelligible that such mechanisms can deliver the rich and varied grip upon the world that we humans enjoy. This, of course, is exactly what PP sets out to achieve.

2 Structural coupling and the bringing forth of worlds¹

Madary notes, more or less in passing, that the PP vision of "organism-relative perceptual content" bears a close resemblance to views that have been defended under the broad banner of "enactivism". I want to pick up on this hint, and suggest that the PP account actually sets the scene for peace to be declared between the once-warring camps of representationalism and enactivism. Thus consider the mysterious-sounding notion of "enacting a world", as that notion appears in [Varela et al. \(1991\)](#)². [Varela et al.](#) write that:

The overall concern of an enactive approach to perception is not to determine how some perceiver-independent world is to be recovered; it is, rather, to determine the common principles or lawful linkages between sensory and motor systems that explain how

1 Parts of this section condense and draw upon materials from [Clark \(in press\)](#).

2 There is now a large, and not altogether unified, literature on enaction. For our purposes, however, it will suffice to consider only the classic statement by [Varela et al. \(1991\)](#). Important contributions to the larger space of enactivist, and enactivist-inspired, theorizing include [Noë \(2004, 2010, this collection\)](#), [Thompson \(2010\)](#), and [Froese & Di Paolo \(2011\)](#). The edited volume by [Stewart et al. \(2010\)](#) provides an excellent window onto much of this larger space.

action can be perceptually-guided in a perceiver-dependent world. (1991, p. 173)

This kind of relation is described by Varela et al. as one of “structural coupling” in which “the species brings forth and specifies its own domain of problems” (1991, p. 198) and in that sense “enacts” or brings forth (1991, p. 205) its own world. In discussing these matters, Varela et al. are also concerned to stress that the relevant histories of structural coupling may select what they describe as “non-optimal” features, traits, and behaviors: ones that involve “satisficing” (see Simon 1956) where that means settling for whatever “good enough” solution or structure “has sufficient integrity to persist” (Varela et al. 1991, p. 196). PP, I will now suggest, has the resources to cash these enactivist cheques, depicting the organism and the organism-salient world as bound together in a process of mutual specification in which the simplest approximations apt to support a history of viable interaction are the ones that are learnt, selected, and maintained.

The simplest way in which a PP-style organism might be said to actively construct its world is by sampling. Action, as Madary noted, serves perception by moving the body and sense-organs around in ways that aim to “serve up” predicted sequences of high-reliability, task-relevant information. In this way, different organisms and individuals may selectively sample in ways that both actively construct and continuously confirm the existence of different “worlds”. It is in this sense that, as Friston, Adams, and Montague (Friston et al. 2012, p. 22) comment, our implicit and explicit models might be said to “create their own data”.³ Fur-

thermore, the PP framework depicts perception and action as a single (neurally distributed) process whose goal is the reduction of salient prediction-error. To be sure, “sensory” and “motor” systems specialize in different predictions. But the old image of sensory information IN and motor output OUT is here abandoned. Instead, there is a unified sensorimotor system aiming to predict the full range of sensory inputs – inputs that are often at least partially self-selected and that include exteroceptive, proprioceptive (action-determining), and interoceptive elements. Nor is it just the sensorimotor system that is here in play. Instead, the whole embodied organism (as Madary notes) is treated as a prediction-error minimizing device.

The task of the generative model in all these settings is (as noted in Clark this collection) to capture the simplest approximations that will support the actions required to do the job. And that means taking into account whatever work can be done by a creature’s morphology, physical actions, and socio-technological surroundings. Such approximations are constrained to “provide the simplest (most parsimonious) explanations for sampled outcomes” (Friston et al. 2012, p. 22). This respects the enactivist’s stress on biological frugality, satisficing, and the ubiquity of simple but adequate solutions that make the most of brain, body, and world. At this point, all the positive enactivist cheques mentioned above have been cashed.

But one outstanding debt remains. To broker real and lasting peace, we must tiptoe bravely back into some muddy and contested territory: the smoking battleground of the Representation wars.

3 Representations: What are they good for?

PP, Madary suggests, provides a new kind of lever for naturalizing intentionality and mental content. Might it also offer a new perspective upon the vexed topic of internal representation? Varela et al. are explicit that, on the enactivist conception “cognition is no longer seen as problem solving on the basis of representations”

³ Such a process repeats at several organizational scales. Thus we humans do not merely sample some natural environment. We also structure that environment by building material artifacts (from homes to highways), creating cultural practices and institutions, and trading in all manner of symbolic and notational props, aids, and scaffoldings. Some of our practices and institutions are also designed to *train us to sample* our human-built environment more effectively – examples would include sports practice, training in the use of specific tools and software, learning to speed-read, and many, many more. Finally, some of our technological infrastructure is now self-altering in ways that are designed to reduce the load on the predictive agent, learning from our past behaviors and searches so as to serve up the right options at the right time. In all these ways, and at all these interacting scales of space and time, we build and selectively sample the very worlds that – in iterated bouts of statistically-sensitive interaction – install the generative models that we bring to bear upon them.

(1991, p. 205). PP, however, deals extensively in internal models – models that may (see [Clark this collection](#)) be rich, frugal, and all points in-between. The role of such models is to control action by predicting and bringing about complex plays of sensory data. This, the enactivist might fear, is where our promising story about neural processing goes conceptually astray. Why not simply ditch the talk of inner models and internal representations and stay on the true path of enactivist virtue?

This issue requires a lot more discussion than I can attempt here.⁴ Nonetheless, the remaining distance between PP and the enactivist may not be as great as that bald opposition suggests. We can begin by reminding ourselves that PP, although it openly trades in talk of inner models and representations, invokes representations that are action-oriented through and through. These are representations that are fundamentally in the business of serving up actions within the context of rolling sensorimotor cycles. Such representations aim to *engage* the world, rather than to depict it in some action-neutral fashion, and they are firmly rooted in the history of organism-environment interactions that served up the sensory stimulations that installed the probabilistic generative model. What is on offer is thus just about maximally distant from a passive (“mirror of nature” – see [Rorty 1979](#)) story about the possible fit between model and world. For the test of a good model is how well it enables the organism to engage the world in a rolling cycle of actions that maintain it within a window of viability. The better the engagements, the lower the information-theoretic free energy (this is intuitive, since more of the system’s resources are being put to “effective work” in engaging the world). Prediction error reports this information-theoretic free energy, which is mathematically constructed so as always to be greater than “surprisal” (where this names the sub-personally computed implausibility of some sensory state given a model of the world – see [Tribus 1961](#)). Notice also that the prediction task uses only information *clearly*

available to the organism, and is ultimately defined over the energies that impinge on the organism’s sensory surfaces. But finding the best ways to predict those energetic impacts can (as substantial bodies of work in machine learning amply demonstrate⁵) yield a structured grip upon a world of interacting causes.

This notion of a *structured* grip is important. Early connectionist networks were famously challenged ([Fodor & Pylyshyn 1988](#)) by the need to deal with structure – they were unable to capture part-whole hierarchies, or complex nested structures in which larger wholes embed smaller components, each of which may itself be some kind of structured entity. For example, a city scene may consist of a street populated by shops and cars and people, each of which is also a structured whole in its own right. Classical approaches benefitted from an easy way of dealing with such issues. There, digital objects (symbol strings) could be composed of other symbols, and equipped with pointers to further bodies of information. This apparatus was (and remains) extremely biologically suspect, but it enabled nesting, sharing, and recombination on a grand scale – see [Hinton \(1990\)](#) for discussion. Such systems could easily capture structured (nested, often hierarchical) relationships in a manner that allowed for easy sharing and recombination of elements. But they proved brittle and inflexible in other ways, failing to display fluid context-sensitive responsiveness, and floundering when required to guide behavior in time-pressured real-world settings.⁶

Connectionist research has since spawned a variety of methods – some more successful than others – for dealing with structure in various domains. At the same time, work in robotics and in embodied and situated cognitive science has explored the many ways in which structure in the environment (including the highly structured artificial environments of text and external symbol systems) could be exploited so as to reap some of the benefits associated with classical forms of in-

⁴ I have engaged such arguments at length elsewhere – see [Clark \(1989, 1997, 2008, 2012\)](#). For sustained arguments *against* the explanatory appeal to internal representation, see [Ramsey \(2007\)](#), [Chemero \(2009\)](#), [Hutto & Myin \(2013\)](#). For some useful discussion, see [Sprevak \(2010, 2013\)](#), [Gallagher et al. \(2013\)](#).

⁵ For reviews and discussions, see [Bengio \(2009\)](#), [Huang & Rao \(2011\)](#), [Hinton \(2007\)](#), and [Clark \(in press\)](#).

⁶ For a sustained discussion of these failings, and the attractions of connectionist (and post-connectionist) alternatives, see [Clark \(1989, 1993, 2014\)](#), [Bechtel & Abrahamsen \(2002\)](#), [Pfeifer & Bongard \(2007\)](#).

ner encoding, without (it was hoped) the associated costs of biological implausibility – see, for example, Pfeifer & Bongard (2007). Perhaps the combination of a few technical patches and a much richer reliance upon the use of structured external resources would address the worries about dealing with structure? Such was the hope of many, myself included.

On this project, the jury is still out. But PP can embrace these insights and economies while providing a more powerful overall solution. For it offers a biologically plausible means, consistent (we saw) with as much reliance on external scaffolding as possible, of internally encoding and deploying richly structured bodies of information. This is because each PP level (perhaps these correspond to cortical columns – this is an open question) treats activity at the level below as if it were sensory data, and learns compressed methods to predict those unfolding patterns. This results in a very natural extraction of nested structure in the causes of the input signal, as different levels are progressively exposed to different re-codings, and re-re-codings of the original sensory information. These re-re-codings (I think of them as representational re-descriptions in much the sense of Karmiloff-Smith 1992) enable us, as agents, to lock us onto worldly causes that are ever more recondite, capturing regularities visible only in patterns spread far in space and time. Patterns such as weather fronts, persons, elections, marriages, promises, and soccer games. Such patterns are the stuff of which human lives, and human mental lives, are made. What locks the *agent* on to these familiar patterns is, however, the whole multi-level processing device (sometimes, it is the whole machine in action). That machine works (if PP is correct) because each level is driven to try to find a compressed way to predict activity at the level below, all the way out to the sensory peripheries. These nested compressions, discovered and annealed in the furnace of action, are what I (following Hinton 1990) would like to call “internal representations”.

What are the *contents* of the many states governed by the resulting structured, multi-level, action-oriented probabilistic generative models? The generative model issues predictions that estimate various identifiable worldly states (includ-

ing states of the body, and the mental states of other agents).⁷ But it is also necessary, as we saw in Clark (this collection) to estimate the context-variable reliability (precision) of the neural estimations themselves. It is these precision-weighted estimates that drive action, and it is action that then samples the scene, delivering percepts that select more actions. Such looping complexities exacerbate an important consequence that Madary nicely notes. They make it even harder (perhaps impossible) adequately to capture the contents or the cognitive roles of many key inner states and processes using the terms and vocabulary of ordinary daily speech. That vocabulary is “designed” for communication, and (perhaps) for various forms of cognitive self-stimulation (see Clark 2008). The probabilistic generative model, by contrast, is designed to engage the world in rolling, uncertainty-modulated, cycles of perception and action. Nonetheless, high-level states of the generative model will target large-scale, increasingly invariant patterns in space and time, corresponding to (and allowing us to keep track of) specific individuals, properties, and events despite large moment-by-moment variations in the stream of sensory stimulation. Unpacked via cascades of descending prediction, such higher-level states simultaneously inform both perception and action, locking them into continuous circular causal flows. Instead of simply describing “how the world is”, these models – even when considered at those “higher” more abstract levels – are geared to engaging those aspects of the world that matter to us. They are delivering a grip on the *patterns that matter* for the *interactions that matter*.

Could we perhaps (especially given the likely difficulties in specifying intermediate-level contents in natural-language terms) have told our story in entirely non-representational terms, without invoking the concept of a hierarchical probabilistic generative *model* at all? One should always beware of sweeping assertions about what might, one day, be explanatorily possible! But as things stand, I simply don’t see how this is to be achieved. For it is surely that very model-invoking

⁷ Bayesian perceptual and sensorimotor psychology (see for example, Rescorla 2013; Körding & Wolpert 2006) already has much to say about just what worldly and bodily states these may be.

schema that allows us to understand how it is that these looping dynamical regimes arise and enable such spectacular results. The regimes arise and succeed because the system self-organizes around prediction-error so as to capture organism-salient patterns, at various scales of space and time, in the (partially self-created) input stream. These patterns specify complex, inter-animated structures of bodily and worldly causes. Subtract this guiding vision and what remains is just the picture of complex looping dynamics spanning brain, body, and world. Consider those same looping dynamics from the multi-level model-invoking explanatory perspective afforded by PP, however, and many things fall naturally into place. We see how statistically-driven learning can unearth interacting distal and bodily causes in the first place, revealing a structured world of human-sized opportunities for action; we see why, and exactly how, perception and action can be co-constructed and co-determining; and we unravel the precise (and happily un-mysterious) sense in which organisms may be said to bring forth their worlds.

4 Predicting peace: An end to the war over internal representation

Dynamically speaking, the whole embodied, active system here self-organizes around the organismically-computable quantity “prediction error”. This is what delivers that multi-level, multi-area, grip on the evolving sensory barrage – a grip that must span multiple spatial and temporal scales. Such a grip simultaneously determines perception and action, and it selects (enacts) the ongoing stream of sensory bombardment itself. The generative model that here issues sensory predictions is thus nothing but that multi-level, multi-area⁸, multi-scale, body-and-action involving grip on the unfolding sensory stream. To achieve that grip is

⁸ The point about multiple areas (not just multiple levels within areas) is important, but it is often overlooked in philosophical discussions of predictive processing. Different neural areas are best-suited – by location, inputs, structure, and/or cell-type – to different kinds of prediction. So the same overarching PP strategy will yield a complex economy in which higher-levels predict lower levels, but different areas learn to trade in very different kinds of prediction. This adds great dynamical complexity to the picture, and requires some means for sculpting the flow of information among areas. I touch on these issues in Clark (this collection). But for a much fuller exploration, see Clark (in press).

to know the structured and meaningful world that we encounter in experience and action.

Is this an inner economy bloated with representations, detached from the world? Not at all. This is an inner economy geared for action, whose inner states bear contents in virtue of the way they lock embodied agents onto properties and features of their worlds. But it is simultaneously a structured economy built of nested systems, whose communal project is both to model and engage the (organism-relative) world.

References

- Bechtel, W. & Abrahamsen, A. (2002). *Connectionism and the mind: Parallel processing, dynamics, and evolution in networks*. Oxford, UK: Basil Blackwell.
- Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2 (1), 1-127. [10.1561/22000000006](https://doi.org/10.1561/22000000006)
- Chemero, A. (2009). *Radical embodied cognitive science*. Cambridge, MA: MIT Press.
- Clark, A. (1989). *Microcognition: Philosophy, cognitive science and parallel distributed processing*. Cambridge, MA: MIT Press.
- (1993). *Associative engines: Connectionism, concepts and representational change*. Cambridge, MA: MIT Press.
- (1997). *Being there: Putting brain, body and world together again*. Cambridge, MA: MIT Press.
- (2008). *Supersizing the mind: Action, embodiment, and cognitive extension*. New York, NY: Oxford University Press.
- (2012). Dreaming the whole cat: Generative models, predictive processing, and the enactivist conception of perceptual experience. *Mind*, 121 (483), 753-771. [10.1093/mind/fzs106](https://doi.org/10.1093/mind/fzs106).
- (2014). *Mindware: An introduction to the philosophy of cognitive science*. New York, NY: Oxford University Press.
- (2015). Embodied prediction. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Clark, A. (in press). *Surfing uncertainty: Prediction, action, and the embodied mind*. New York, NY: Oxford University Press.
- Fodor, J. & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28 (1-2), 3-71. [10.1016/0010-0277\(88\)90031-5](https://doi.org/10.1016/0010-0277(88)90031-5)

- Friston, K., Adams, R. & Montague, R. (2012). What is value—Accumulated reward or evidence? *Frontiers in Neurobotics*, 6. [10.3389/fnbot.2012.00011](https://doi.org/10.3389/fnbot.2012.00011)
- Froese, T. & Di Paolo, E. A. (2011). The enactive approach: Theoretical sketches from cell to society. *Pragmatics and Cognition*, 19 (1), 1-36. [10.1075/pc.19.1.01-fro](https://doi.org/10.1075/pc.19.1.01-fro)
- Gallagher, S., Hutto, D., Slaby, J. & Cole, J. (2013). The brain as part of an enactive system. *Behavioral and Brain Sciences*, 36 (4), 421-422. [10.1017/S0140525X12002105](https://doi.org/10.1017/S0140525X12002105)
- Hinton, G. E. (1990). Mapping part-whole hierarchies into connectionist networks. *Artificial Intelligence*, 46 (1-2), 47-75. [10.1016/0004-3702\(90\)90004-J](https://doi.org/10.1016/0004-3702(90)90004-J)
- (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences*, 11 (10), 428-434. [10.1016/j.tics.2007.09.004](https://doi.org/10.1016/j.tics.2007.09.004)
- Huang, Y. & Rao, R. (2011). Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2 (5), 580-593. [10.1002/wcs.142](https://doi.org/10.1002/wcs.142)
- Hutto, D. D. & Myin, E. (2013). *Radicalizing enactivism: Basic minds without content*. Cambridge, MA: MIT Press.
- Karmiloff-Smith, A. (1992). *Beyond modularity: A developmental perspective on cognitive science*. Cambridge, MA: MIT Press/Bradford Books.
- Körding, K. & Wolpert, D. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences*, 10 (7), 319-326. [0.1016/j.tics.2006.05.003](https://doi.org/10.1016/j.tics.2006.05.003)
- Madary, M. (2015). Extending the explanandum for predictive processing - A commentary on Andy Clark. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Noë, A. (2004). *Action in perception*. Cambridge, MA: MIT Press.
- (2010). *Out of our heads: Why you are not your brain, and other lessons from the biology of consciousness*. New York, NY: Farrar, Straus and Giroux.
- (2015). Concept pluralism, direct perception, and the fragility of presence. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Pfeifer, R. & Bongard, J. (2007). *How the body shapes the way we think: A new view of intelligence*. Cambridge, MA: MIT Press.
- Ramsey, W. M. (2007). *Representation reconsidered*. Cambridge, UK: Cambridge University Press.
- Rescorla, M. (2013). Bayesian perceptual psychology. *Oxford handbook of the philosophy of perception (forthcoming)*. Oxford, UK: Oxford University Press.
- Rorty, R. (1979). *Philosophy and the mirror of nature*. Princeton, NJ: Princeton University Press.
- Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, 63 (2), 129-138. [10.1037/h0042769](https://doi.org/10.1037/h0042769)
- Sprevak, M. (2010). Computation, individuation, and the received view on representation. *Studies in History and Philosophy of Science*, 41 (3), 260-270. [10.1016/j.shpsa.2010.07.008](https://doi.org/10.1016/j.shpsa.2010.07.008)
- (2013). Fictionalism about neural representations. *The Monist*, 96 (4), 539-560. [10.5840/monist201396425](https://doi.org/10.5840/monist201396425)
- Stewart, J., Gapenne, O. & Di Paolo, E. (Eds.) (2010). *Enaction: Towards a new paradigm for cognitive science*. Cambridge, MA: MIT Press.
- Thompson, E. (2010). *Mind in life: Biology, phenomenology, and the sciences of mind*. Cambridge, MA: Harvard University Press.
- Tribus, M. (1961). *Thermodynamics and thermostatics: An introduction to energy, information and states of matter, with engineering applications*. New York, NY: D. Van Nostrand Company Inc.
- Varela, F., Thompson, E. & Rosch, E. (1991). *The embodied mind*. Cambridge, MA: MIT Press.

Levels

Carl F. Craver

The levels metaphor is commonly used to describe science, its theories, and the world. Yet the metaphor means different things in different contexts, inviting equivocation. These distinct applications of the metaphor can be distinguished by the relata they relate, the relation between levels that they assert, and the rule by which they locate items at a level. I discuss these many applications of the levels metaphor with an eye to developing a descriptively adequate account of one particular application: levels of mechanisms. I argue that this application of the metaphor is central to the explanatory practices of the special sciences and defensible as a metaphysical picture of how phenomena studied in the special sciences are constituted.

Keywords

Emergence | Explanation | Hierarchy | Interlevel causation | Mereology | Reduction

Author

Carl F. Craver

ccraver@artsci.wustl.edu

Washington University
St. Louis, MO, U.S.A.

Commentator

Denis C. Martin

denis.martin@hu-berlin.de

Humboldt Universität zu Berlin
Berlin, Germany

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

The levels metaphor is ubiquitous in our descriptions of science and the world. So simple and elegant, the metaphor takes an apparently heterogeneous collection of objects and arranges them in space from bottom to top. The metaphor works in so many contexts because it leaves open just what kinds of object are to be arranged, what distinguishes top from bottom, and what it means to say that an object is at some levels and not others. This flexibility explains the metaphor's fecundity, but it also helps to obscure the fact that it is used in many different ways in many different contexts.

A survey of kinds of levels drawn from science and philosophy would have to include levels of abstraction (Floridi 2008), aggregation (Wimsatt 1997), analysis (Shepherd 1994;

Churchland & Sejnowski 1992), causation and explanation (Kim 1998), implementation (Marr 1982), organization (Churchland & Sejnowski 1992), processing (Craik & Lockhart 1972), realization (Gillett 2002), sizes (Wimsatt 1976), sciences, theories, and explanations (Oppenheim & Putnam 1958). Many of these familiar applications of the levels metaphor are distinct from but also clearly related to one another. And when they are related, they often have rather indirect and reticulate connections. The level metaphors thus takes subtly different forms when applied in neighboring contexts, and this obscures the extent to which features of one application of the metaphor do and do not transfer from one context to the next. My first thesis, then, is that our ways of describing sci-

ence and the world contain many distinct, legitimate applications of the levels metaphor that are either unrelated or that have only indirect relations with one another.¹

This *descriptive pluralism* about the levels metaphor is directly opposed to eliminativism about levels (Fehr 2004; Machamer & Sullivan 2001; Thalos 2013). The suggestion that we might be better off abandoning the levels metaphor is about as likely to win converts as the suggestion that we should abandon metaphors involving weight or spatial inclusion. These metaphors are too basic to how we organize the world to seriously recommend that they could or should be stricken from thought and expression. Yet, descriptive pluralism about the levels metaphor is consistent with the thought that some applications of the metaphor distort the structure of the world or represent it as having conceptually incoherent structures. I discuss some examples below. The central message of this paper is that there can be no single verdict concerning the utility or conceptual soundness of the levels metaphor *simpliciter*. The metaphor must be evaluated and used with caution, especially when it is called on to settle disputes about the character of science and the metaphysical structure of the world.

As some motivation for adopting this proposal, and as a step toward a more positive thesis, I show that we can avoid some simple confusion by separating the different applications of the metaphor. To make this case, I build slowly toward a particular application of the metaphor that, as I have argued elsewhere (Craver 2001, 2007), is central to explanatory practices in neuroscience and across the special sciences: levels of mechanisms. This application of the levels metaphor is metaphysically plausible and, so far as I can tell, more or less innocuous; that is part of its virtue. Yet this simple and useful application of the metaphor can begin to appear problematic when it is inappropriately assimilated to other applications that

serve altogether different purposes in our thinking about science and the world.

My point is not to defend levels of mechanisms as the one true application of the levels metaphor (that would be as pointless as eliminativism). Rather, my first positive goal is to provide a reasonably clear account of levels of mechanisms and to show that this application is metaphysically benign yet exceptionally important for doing science. Levels of mechanisms are, as would be expected, richly but indirectly connected with many other applications of the metaphor. My second goal is to highlight and disentangle some of the confusions that arise from failing to keep levels of mechanisms distinct from other senses of levels. In particular, I show that commitment to the existence of levels of mechanisms entails no commitment to: a) monolithic levels in nature, b) the stratification of sciences by levels, or c) a tidy hierarchy of theories among the sciences. I will also show why levels of mechanisms are d) distinct from Marr's views about levels of abstraction and e) distinct from levels of realization more generally. I argue that f) the idea of interlevel causation is conceptually awkward within levels of mechanisms (but not to levels of size, for example). Furthermore, g) the idea of levels of mechanisms nicely expresses the idea of emergence as a kind of non-aggregativity while providing no support to those who seek evidence in biology for a more robust kind of emergence. The failure to disambiguate altogether separate applications of the levels metaphor creates a conceptual malaise for which levels of mechanisms are at least a partial cure.

2 Refining the levels metaphor: Three defining questions

In its barest of forms, the levels metaphor demands little of its object; it requires only a set of items and some criterion for ranking them as higher or lower than one another in some respect. Seniors are at a higher level in the American high school system than juniors, poetry is at a higher level than pushpin, lust is at a higher (and lower) level than like, and cells are at a higher level than molecules. In these ex-

¹ Standard etymologies trace the term “level” to the balance and from there to the idea of a flat, horizontal landing, as in the stories of a building. From there, it is easy to see how the metaphor might be extended to the kinds of hierarchy discussed in this paper.

amples, it is obvious that different kinds of thing are related by entirely distinct kinds of relation. In subtler cases, the equivocation is less noticeable, and for that, all the more misleading.

Three defining questions can be used to explicate how the levels metaphor is applied in a given context:

The Relata Question: What kinds of item are being sorted into levels?

The Relations Question: In virtue of what are two items at different levels?

The Placement Question: In virtue of what are two items at the same level?

The Relata Question provides an important clue about the intended sense of levels. The flexibility of the metaphor allows it to be applied to *abstracta*, such as branches of mathematics and ethical principles, or to *concreta*, such as astronomical objects and stereo equipment. The metaphor can be applied to types, such as sergeants and corporals, or to tokens, such as the relationship between Colonel Blake and Corporal O'Reilly. It can be applied to objects such as cats and mountains, to activities such as releasing neurotransmitters and making decisions, and to properties such as excitability or charge. Within the neurosciences, the levels metaphor is applied fluidly to causes, descriptions, developmental stages, events, explanations, scientific fields, objects, properties, techniques, and theories. Confusion arises when we assume that each application is the same.

The Relations Question concerns the ordering relationship by which items are said to be at a higher or a lower level than one another. A theory, for example, might be said to be at a higher level than a second if the first is derivable from the second (and not vice versa); the lowest-level theories are in this sense “fundamental.” Poetry might be said to be higher than pushpin in the sense that it requires greater intellectual skill and training to take pleasure in the former than to take pleasure in the latter. A technique might be said to be at a lower level than another because it detects phenomena at a smaller size scale. Some applications of the levels metaphor are discrete in the sense that there is a gap between things at lower and high

levels; other applications are continuous, as when one uses the metaphor to describe size. We are unlikely to confuse such wildly different kinds of relationship. However, as we will see, the metaphor is used in other contexts where it is beguilingly difficult to keep them distinct, even for those who know better.

The Placement Question asks for the principle by which different items are located on the same level. Many uses of the levels metaphor rely at heart on an answer to the placement question. When the metaphor is used to describe size scales, for example, puffins and porcupines are at roughly the same level, vasopressin and oxytocin are at roughly the same level, and hydrogen and oxygen atoms are together at a lower level still. Juniors are all juniors because they are in their third year of American high school. For Marr, computational level questions are directed at what is computed and why it is computed that way (Shagrir 2010; Bechtel & Shagrir 2013).² Not every account of levels requires an answer to the placement question affirmatively. Indeed, it is of central importance that the idea of levels of mechanisms articulated here entails no positive story about what it means to be at a level, only a negative story about when things are not at different levels.

3 From gesture to prototype

Perhaps the most common application of the levels metaphor is to gesture loosely at the relationship between different fields of scientific research, *levels of science*.³ In neuroscience, for example, some researchers work at “the molecular level,” doing things such as sequencing channel proteins, studying enzyme kinetics, or manipulating genes. Others work at the cellular level, doing things such as staining cells, recording action potentials, or studying neural migration. Others study brain regions, characterizing

² One consequence of the following discussion is that not every account of levels must offer a unique answer to the placement question. Levels of mechanisms are defined by their distinctive relata and relations; these constraints, by themselves, offer no unique answer to the placement question. This is why levels of mechanisms are, as I will argue, local rather than monolithic.

³ For the relevant sense of a scientific “field”, see Darden & Maul (1977) and Darden (1992).

their anatomical features or, studying the propagation of neural signals within them. Still others work at the level of systems, using functional magnetic resonance imaging (fMRI), transcranial magnetic stimulation (TMS), and cognitive tasks to find large-scale cognitive systems in the mind-brain. One could perhaps insinuate other levels between these, and one could certainly extend the hierarchy further down or higher up. But the central idea is that the *scientific fields* can be ordered as higher or lower than one another.

Scientific fields are individuated in part by their theories (Darden 1992). The gestural sense of levels, then, can seem to carry the implication that scientific *theories* are or will someday be ordered more or less clearly into levels. Oppenheim & Putnam's (1958) influential view of the unity of science is based on a rough correspondence between levels of science, levels of theory, and levels of parts and wholes (see Table 1). They divide the world into six ontological strata (societies, organisms, cells, molecules, atoms, and elementary particles). These strata are defined by mereological relationships among types: elementary particles are the parts of atoms, atoms are the parts of molecules, molecules are parts of cells, and so on. Each of these strata is assigned a distinct science: economics and the social sciences at the top, particle physics at the bottom. Each science develops its theory more or less autonomously from the others, so the theories developed by these sciences can themselves be ordered, like the layers of a cake, from top to bottom. The unity of science, for Oppenheim and Putnam, is to be achieved by explaining phenomena in the domain of a higher-level science, as described in the theory of that science, in terms of the items in the domain of the more fundamental science, as described in the theories of that science. (Levels of mechanisms, as defined below, involve a kind of part-whole relation as well but without any commitment to the idea that such type-level part-whole relationships correspond in even a rough way to the structure of the sciences or to the structures of their theories.)

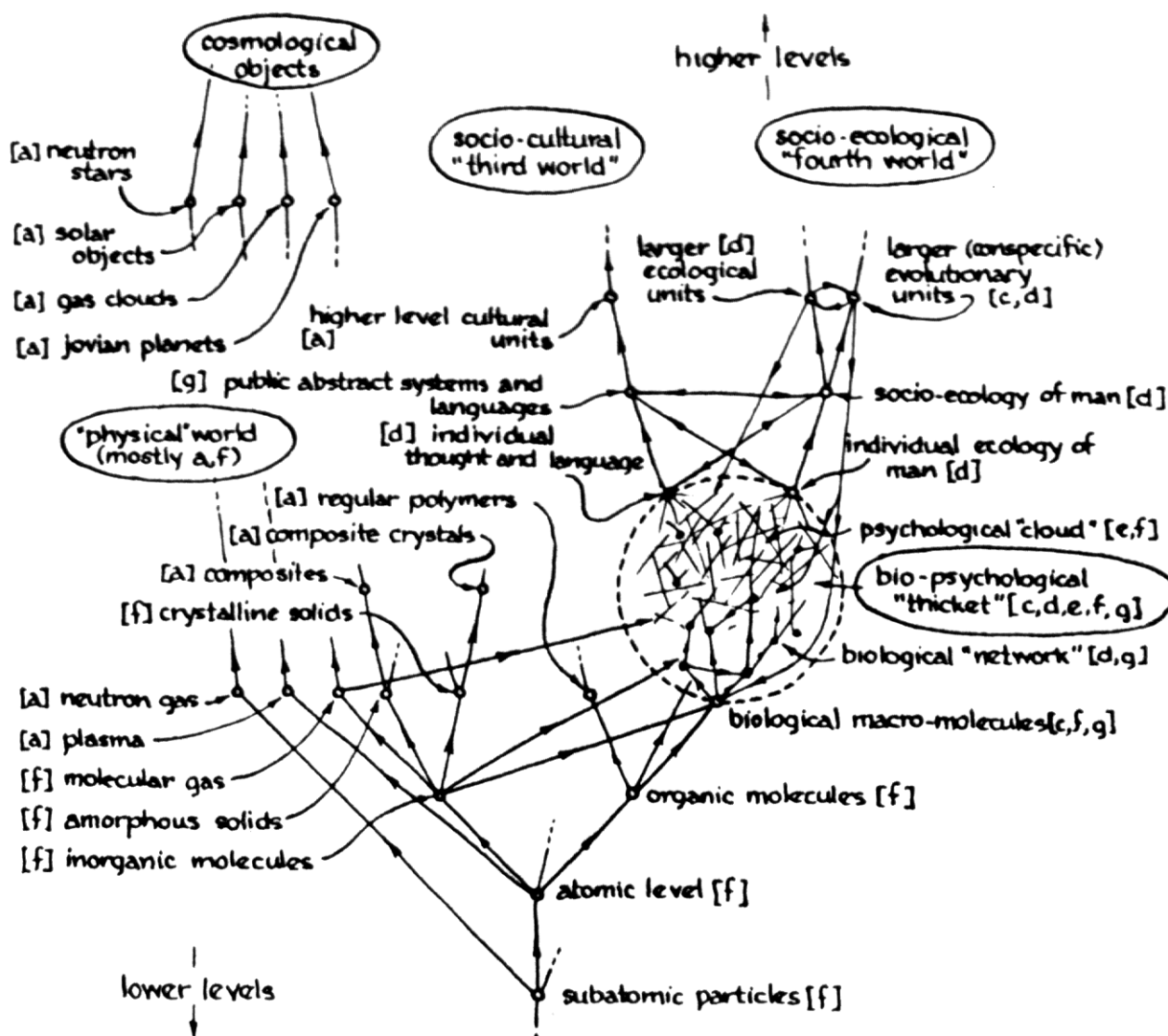
Wimsatt's detailed and influential exploration of the levels metaphor confronts Oppen-

heim and Putnam with the complexity of the levels found in many areas of contemporary science (Wimsatt 1976). Against Oppenheim and Putnam's six-layer model, Wimsatt's "Reductionist Illustrative" (Figure 1) represents multiple branches of levels fanning out from the lowest level in subatomic particles to cosmological objects, the sociocultural world (e.g., economic and political phenomena), and the socioecological world (e.g., evolution).

Wimsatt's tree diagram, however, represents only one aspect of his *prototype account* of levels that encompasses many more features than Oppenheim and Putnam's layer-cake mapping in Table 1. The core features in Wimsatt's prototype are:

- **Size.** Higher-level items are larger than lower-level items.
- **Composition.** Higher-level items are made up of lower-level objects and processes.
- **Laws.** Laws of nature hold only or mostly between items at the same level.
- **Forces.** Distinct forces operate at different levels.
- **Predictability.** Levels are local maxima of regularity and predictability that appear at different size scales.
- **Detection.** Items at a given level tend to be detected or detectable primarily by other items at that level.
- **Causes.** Causal relationships hold only or mostly between items at the same level.
- **Theories.** Scientific theories describe phenomena exclusively or mostly at a single level.
- **Techniques.** Different techniques and instruments detect items at different levels.
- **Disciplines.** Different disciplines of science direct their attention at different levels.

Wimsatt's view is a prototype view in the sense that it characterizes the levels metaphor in terms of a core set of features, not all of which must be present in order for the metaphor to apply. Insofar as Wimsatt embraces a prototype model, he can be seen as embracing descriptive pluralism while, at the same time, holding that there is a sufficiently strong family resemblance



(h) A reductionistic (?) illustrative*
phylogenetic ontology of our world as we see it.

[letters in parentheses refer to local character of network around that node]

* A diagram like this is obviously highly tentative. Although it has been constructed with many specific relations in mind, I would claim accuracy only of a general qualitative sort – i.e., for the rough general distribution of network properties, as indicated, e.g., by the distribution of letters.

W.C. Wimsatt 1973

Figure 1: Wimsatt's tree of levels branches to preserve compositional relationships among levels (1976).

among the plurality of applications of the levels metaphor to warrant their inclusion in a single prototype.

Is the levels metaphor sufficiently unified across these different applications to warrant a single prototype? My remarks on the relation, and placement questions should already indicate that it is not—that different features in Wimsatt's list are at best indirectly related and

so fail to map to one another in any tidy way. While the prototype approach usefully highlights the complexity of the levels metaphor, it also obscures the extent to which the different features in the prototype are features of different applications of that metaphor.⁴

⁴ Wimsatt's diagram in Figure 1 reflects this. The branching tree structure is ordered by compositional relations. Wimsatt's view of levels as dissipating waves (see Figure 3 below) flouts that relation.

Table 1: Oppenheim and Putnam's layer-cake sketch of the levels of mereology (left), sciences (middle), and theories (right).

Mereological level	Sciences	Theories
Societies	Economics	Classical Economics
Organisms	Psychology	Law of Effect
Cells	Cytology	The Neuron Doctrine
Molecules	Chemistry	The Central Dogma
Atoms	Physics	The Bohr Model
Sub-Atomic Particles	Quantum Mechanics	Schrödinger Equation

4 Levels of sciences and theories

Wimsatt, Oppenheim, and Putnam all include within their analysis of levels the gestural idea that different fields or disciplines of science are arranged by the sizes of the objects they study. Wimsatt's branching hierarchies depict a more ornate structure. Within that structure, it seems inappropriate to say that astrophysics is at a higher level than biology or economics, though astrophysicists typically deal with things that are orders of magnitude larger than the things biologists and economists study. The gestural sense of levels doesn't seem to branch that way.

When we apply the levels metaphor to *sciences*, the relata are units of scientific organization (such as fields, research programs, or disciplines). Answers to the relations and placement questions are more difficult to discern and are likely impossible to express both accurately and concisely for this application. Size seems to be relevant, but we have just seen that it cannot be the whole story. The branches in Wimsatt's diagram follow, in addition to size relationships, relationships of composition. The things studied by economists (groups) are composed of things studied

by psychologists (organisms), which are composed of things studied by physiologists (physiological systems), and so on. The things studied by Darwin are composed of the things studied by zoologists, which are composed of the things studied by cytologists. The point of these examples is not to get the branches in Wimsatt's hierarchy exactly right; any proposed hierarchy of the sciences and the items in their domains is bound to be historically contingent and provisional at best.

In fact, many sciences appear to resist tidy compartmentalization within levels. Neuroscience, especially cognitive neuroscience, is a paradigm of multilevel science, encompassing the study of ions, ion channels, cells, populations of cells, brain regions, and behaviors of whole organisms. No competent evolutionary biologist can avoid knowing something about genes, physiological systems, organisms, populations, and environments. Many sciences, in short, contain items within their domain that stand in compositional relations to one another. Such sciences often construct multilevel theories that integrate findings across multiple levels of organization. This is one reason why the relationship between levels of science and part-whole levels is indirect.

Another reason is that, in many cases, more than one science is dedicated to studying items at the same mereological or size level. Cytologists, anatomists, and electrophysiologists all study aspects of cells with different tools. The ethologist and the experimental psychologist study animal behavior, but they approach that behavior with different assumptions, methods, and theories. Economists, ecologists, epidemiologists, and organizational psychologists study populations of organisms. The relationship between levels of science and the ontological levels that Oppenheim and Putnam presume is many to many.

For this reason, it is unlikely that any precise answer to the placement question will correctly express the application of the levels metaphor to sciences. One might say that two sciences are on the same level when they pertain to items at the same compositional level. Perhaps it makes sense to say that Camillo Golgi was investigating the same level when he stained Purkinje cells with silver nitrate that Alan Hodgkin was investigating when he used his voltage clamp to study the action potential of the squid giant axon. They were both studying cells, but they studied different phenomena and used different techniques. If we focus now on the parts of these wholes, we see that these different scientists are not even on the same branches of a Wimsatt diagram, and the levels metaphor begins to break down. Ask Golgi about the relevant parts of the cell, and he will tell you about its gross morphological features and its organs. Ask Hodgkin and Huxley about the relevant parts of the squid giant axon, and they will tell you about membranes, axon hillocks, ionic conductances, and voltage gradients. An epidemiologist might talk about nodes and networks and hubs in a model of contagion. Economists will talk about producers and consumers. Differences in scientific interests often entail differences in the relevant ontology for the science; and the same thing can be carved into parts in many ways depending on what one is interested in describing or explaining (Kauffman 1971; Wimsatt 1972).

The take-home lesson: the application of the levels metaphor to fields of science yields a

notion of levels only indirectly related to ontological levels (as understood in a roughly compositional, part-whole sense). The idealized, Oppenheim-Putnam correspondence between levels of science, levels of theory, and levels of mereology breaks down in the face of this many-many mapping. And the compositional aspect in Wimsatt's prototype appears to be only loosely related to the application of the levels metaphor to fields of science. These are, in short, distinct applications of the metaphor, offering different answers to the relata, relations, and placement questions. As a result, an understanding of how sciences can be organized loosely into levels provides no direct insight into ontological levels. This will come as no surprise to those who study intellectual history, or to those who have witnessed for themselves how fields of science change their boundaries over time. Our age, perhaps more than any other, has witnessed an explosion of hybrid fields (neuroeconomics, behavioral genetics, cognitive ethology) that cross levels, combine approaches, and attempt to feed off insights shared between distant scientific neighbors. The historical relativity of disciplinary boundaries makes them unreliable guides to ontology.

The same considerations suggest that *levels of scientific theory* will also have a many-many relationship with ontological levels. In this application of the metaphor, the relata are theories or models. And the relationship is typically construed as a kind of subsumption, e.g., deductive subsumption (Hempel 1965; Schaffner 1993; Kitcher 1989), or some kind of similarity or inclusion (Bickle 1998). The disciplinary hodgepodge of the special sciences fails to match this philosophical reconstruction. Single theories, such as the theory explaining spatial memory in terms of memory systems, grid cell organization, synaptic plasticity, and changes in ionic conductances through a membrane (see Moser 2008), often reach across many different part-whole levels (Darden & Maul 1977; Bechtel 1988; Craver 2002, 2008). One and the same mereological unit (e.g., cells) can appear in many distinct theories (e.g., neurons play some role in most theories in neuroscience).

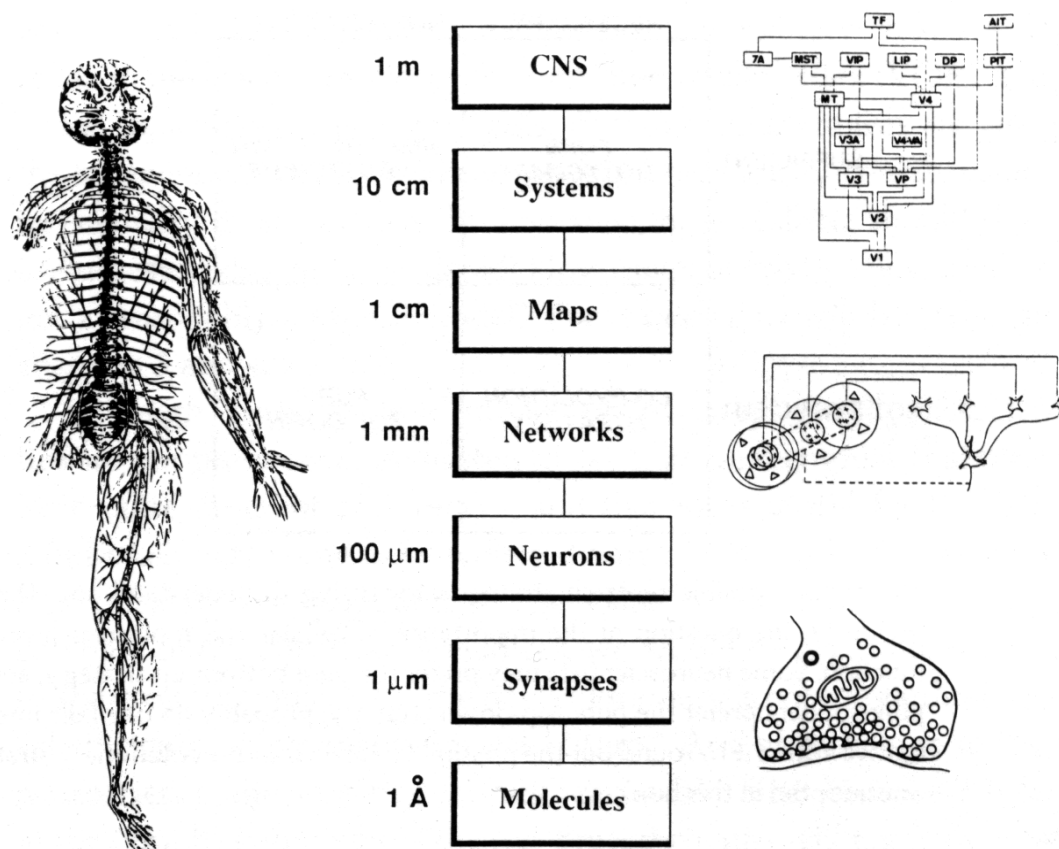


Figure 2: Churchland & Sejnowski's (1992) diagram of levels of organization in the central nervous system.

The multilevel structure of contemporary science emphasized in nearly every corner of the special sciences is not best understood as a hierarchy of theories. Nor is it a hierarchy of fields. Instead, there is an ontological hierarchy working behind the scenes. This background ontological assumption guides the development of theories, informs the criteria for evaluating explanations, and underlies the roughly hewn idea that sciences and theories are organized into levels. It is the expression of an ideal of explanation to understand how things work in terms of their component parts and to understand how those parts work in terms of still lower-level components. It is precisely because the world is presumed to have this kind of multilevel structure, of mechanisms within mechanisms, that the sciences investigating that world and the theories describing it are so reticulate that they can look like the “bio-psychological thicket” on the right side of Wimsatt’s tree. In the thicket, the orderly relationship among

levels breaks down and is replaced by a jumble. The image makes it hard to see any meaningful sense in which distinct items are at different levels. Perhaps this thought fuels eliminativism about levels.

The biological sciences are undeniably thicket-like. But this sociological fact is only indirectly related to the ontic structures presumed to lie behind and scaffold the development of these theories. From now on, then, I focus on applications of the levels metaphor to the world, not to sciences or theories.

5 Size levels

One ontological application of the levels metaphor emphasizes the relative sizes of objects at different levels. The relation in size levels are objects or kinds of object, and the interlevel relationship is relative size (larger, smaller). Things in the same size range are at the same level. Churchland and Sejnowski’s classic diagram of

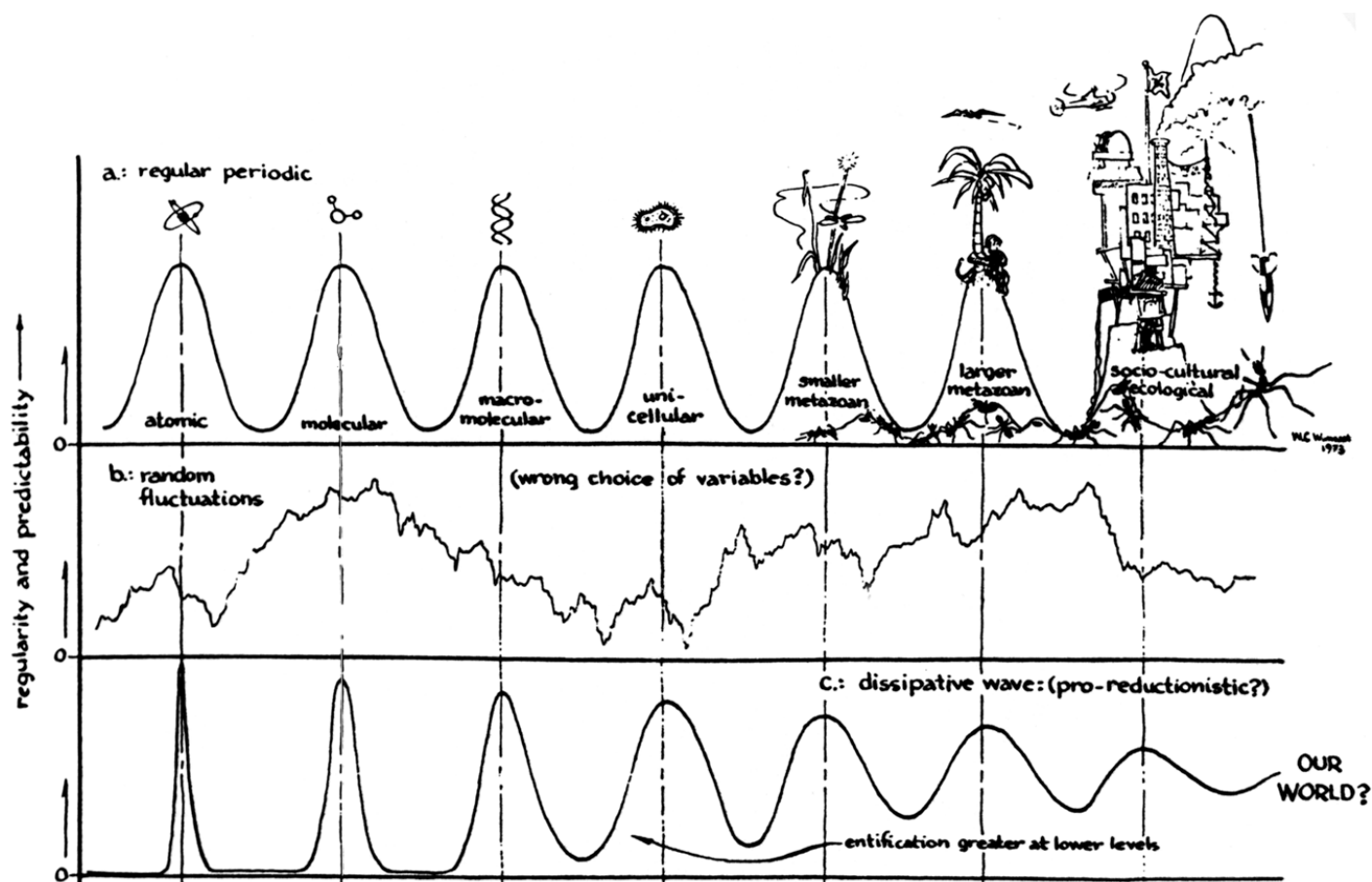


Figure 3: Levels as local maxima of regularity and at different size scales (Wimsatt 1976).

levels in the neurosciences (Figure 2) is accompanied by size scales for each level, ranging from Angstrom units to meters.⁵

As noted above, Wimsatt's tree diagram branches precisely because it is tracking something stronger than size: some kind of compositional relation. In a second diagram (Figure 3), Wimsatt emphasizes size and abandons the compositional relationship implicit in Figure 1. The abscissa in Figure 3 represents a roughly logarithmic size scale, and yet the figure is not compositional. Large metazoan organisms are not generally composed of smaller metazoan organisms, and it would surely be a stretch to claim that these two are generally composed of unicellular organisms (though there might be some truth in that claim). The ordinate in this diagram is a measure of regularity and predictability. The figure is repeated three times, each

illustrating a different way the world might be organized with respect to size. At the top is an orderly world (despite impending doom on the right). Objects become more regular and predictable in their behavior, and to the same degree, at certain size scales. Beneath this is a world with no sharp peaks of regularity and predictability. As Wimsatt notes, scientists confronted by such a world might question whether they have chosen the right variables for their models. On the bottom is Wimsatt's conjecture for our world, where regularity is very high for single atoms but falls off at larger or smaller scales. This wave dissipates over time, peaking lower and spreading out over larger and larger size ranges as scale increases. Wimsatt's diagram thus represents an *empirical hypothesis* about how levels, as peaks of regularity and predictability, are in fact distributed across different size scales in our world. If his empirical hypothesis is correct, it calls out for explanation that our world is more like the first and third graphs than it is like the second.

⁵ One might, analogously, arrange a hierarchy of activities, with different activities occurring on different temporal scales. The idea of levels of mechanisms combines these two ideas; it is a hierarchy of doings framed by a relevance relationship between those at lower levels and those at higher levels.

Why does Wimsatt represent size as the determining factor in regularity and predictability? The answer turns on other features in his prototype account. For example, things of different sizes effect and are affected by different forces, and objects of different sizes act and interact with one another more than they interact with objects at other levels. Market forces run economies, cosmic objects move under gravitational forces, and hydrogen bonds hold molecules together. Regularity and predictability peak at different size scales because the forces act and the causal relationships occur mostly at those size scales.

Wimsatt's empirical hypothesis has not been tested.⁶ Despite its intuitive appeal, one can readily produce examples of causes, forces, and laws that operate promiscuously across a very wide range of size scales. Big things (even very big things) and little things (even very little things) routinely interact, as when planets attract molecules into atmospheres or when a five-millimeter louse attaches itself to a thirteen-meter gray whale. Forces also act at many scales. Gravitation affects the human species on an evolutionary scale just as much as it influences individual human actions and the otoliths in our vestibular system. The very existence of interlevel theories, bridging molecules to behaviors (for example), provides ample evidence that regularity and predictability often span size scales: facts about gasses can be predicted from facts about molecules, and facts about learning can be inferred from facts about molecules.

If we could find a way to test Wimsatt's hypothesis, it might turn out that causes, forces, and laws do tend to cluster around certain size scales. This would be a striking empirical fact about the world and would, again, call out for some kind of explanation. In contrast, Wimsatt raises no *principled* objections to interlevel causes, forces, or regularities; he offers an *empirical* hypothesis that interlevel causes, forces, and regularities tend to be less prevalent than those operating at a single level.

⁶ It is hard to say how it would be tested and, in particular, how predictability is to be measured. Surely items in the valleys of this diagram are not unpredictable, full stop. Rather, they are more difficult to predict for creatures *like us*, unaided by machines and programs. It is not clear why human cognitive abilities should have any further ontological significance.

There are, however, apparent principled, *conceptual* difficulties faced by the effort to describe *levels of realization* in terms of a causal relation. There are many notions of realization, often tailored to altogether distinct philosophical disputes (Craver & Wilson 2007). On most accounts, however, one and the same object or event has both the realized and the realizing property, and the object cannot differ with respect to the realized property without the realizing property being different in some way (supervenience). The relata here are properties. The interlevel relationship is or includes supervenience.⁷

Marr's levels, as I understand them, are levels of realization. The hardware realizes the algorithm, which, in the right context, realizes the computation. It is awkward at best to say that the algorithm causes the computation; rather, the algorithm implements the computation in context. Changing context can change the computation. For example, a subtraction algorithm can implement division; the log of a division is a difference of logs. Likewise, the function represented in the algorithm is not caused by the hardware; the hardware instantiates or implements the algorithm. Computation-, algorithm-, and hardware-level theories are all different ways of describing one and the same thing—different predicates applied to one and the same system as a whole in its working context.

The same holds for what we might call micro-realization: when some property of a whole is realized by the organized and interacting collection of parts that constitute the property of the whole. An early edition of the *Betty Crocker Cookbook* apparently contains an explanation of how the microwave heats the soup (Churchland 1995). According to this explanation, the molecules excited by the microwave rub against one another and heat the soup by friction. As Churchland points out, Betty misrepresents the relationship between the heat of a liquid and the kinetic energy of its constituent molecules. Temperature is not produced by the

⁷ Supplemental conditions might be added to make realization more demanding (e.g., Melnyk 2003; Haug forthcoming). The point I wish to make doesn't turn on this matter.

mean kinetic energy of component molecules in such cases; rather, temperature in such situations is constituted or realized by (Churchland would say identical to) the mean kinetic energy of the components. In the same way, one might think that the behavior of a mechanism as a whole is realized by, rather than caused by, the organized collection of its components. The beating of the heart is realized, not caused, by the choreographed movements of the auricles and ventricles. It is awkward and unnatural to assert otherwise.

The apparent awkwardness and unnaturalness of such ways of talking follows from many core principles that many (rightly or wrongly) embrace about the nature of causation. If one thinks that causes must precede their effects, and one understands the realization relationship as a synchronic relation, then levels of realization cannot be causally related. If one thinks of causation in terms of the intersection of processes and the exchange of marks or conserved quantities, then the relata in levels of realization do not *come to* intersect in space-time (they always and everywhere intersect), they do not carry their marks beyond the locus of the intersection (because they always and everywhere intersect), and they do not pass anything from one to the other. In short, the *intimacy* among levels of realization seemingly precludes any standard metaphor of production, or “oomph,” or expression of a disposition, or the exertion of a power. This intimacy stands in the way of anyone who believes that causes and effects must be altogether distinct from one another.⁸ So indistinct are levels of realization that many philosophers, Churchland included, prefer to speak of identity in such contexts (see Polger 2006). Finally, if one thinks of causation in terms of the ability to manipulate effects by intervening on causes, one will note that there is no way to intervene to change the properties of wholes without, at the same time, intervening to change the supervenience base of those properties.⁹

⁸ I discuss a representative quote from Lewis below when considering causal relations between levels of mechanisms.

⁹ One can, in cases of multiple realization, intervene into the parts and their organization without intervening to change the property of the whole, and this affords some measure of independence. Perhaps one can find room in this view for the idea of understanding bottom-up relations in a hierarchy of realization as causal (though, again, realization or token identity seem to be better ways of talking). But there

I raise these issues not to cement a case against the possibility of understanding realization and causation so as to leave conceptual space for causation between levels of realization. (For a fuller discussion, see Kim 2000; Craver & Bechtel 2006). I mention them only to point out that relations of size and realization have very different implications for the intelligibility of interlevel causation. No theory or principle of causation that I know places any metaphysical restrictions on causal relations among objects of different sizes. Many theories or principles of causation appear to rule out the possibility of causal relationships between levels of realization. The point is that Wimsatt’s empirical hypothesis that causes, laws, and regularities tend to be sequestered within size scales is altogether distinct from the claim that there is no conceptual room for causation between levels of realization. Interlevel causation is mysterious or not depending on which views of levels and causation one adopts.¹⁰

6 Parts and wholes

A distinct and indirectly related application of the levels metaphor in the neighborhood of

is no room in the view (no conceptual room) for causation to work from the top down in such levels. For a penetrating discussion of this matter and its implications for causation in a multilevel world, see Baumgartner 2010, 2013; Romero (forthcoming).

¹⁰ Levels of control and levels of processing, in contrast, are defined in terms of causal relations. In *levels of control*, the relata are agencies and the relation is dominance. Items at higher levels direct or regulate the activities of their underlings. Majors and corporals, queen bees and drones, bosses and workers occupy different levels of a control hierarchy. Analogous relations are sometimes found among physiological systems. When one speaks of “executive function” in cognition, one is describing levels of control.

The idea of control or dominance is a causal notion, and it is independent of matters of size (witness the sauropod brain). Contra Fehr (2004), the idea that the world is organized in levels of realization or organization (as defined below) is not an expression of patriarchy; it is an equivocation to characterize realization and organization as relations of dominance. In levels of control, the relata are logically independent and spatiotemporally distinct interactors. It is not at all implausible for one to control the other causally (more on this below).

In *levels of processing*, the relata are processing units of some sort (such as brain regions or computational modules), and they are related as “upstream” or “downstream” in the flow of information or the order of production. In the early visual system (neglecting feedback for the moment), one can describe visual information as passing from lowest- (shallowest-, earliest-) level processing in the retina to higher- (deeper-, later-) level processing in the Lateral geniculate nucleus (LGN) and the primary visual cortex. Levels of realization and organization are not earlier or later than one another. Craik and Lockhart define levels of processing in terms of depth of semantic or cognitive processing, not in terms of decomposition.

levels of realization invites a different kind of equivocation, this time concerning the existence of higher-level powers. This application involves not a whole-whole relationship but rather the relationship between the behavior or property of a whole and the behaviors or properties of one of its parts. The behavior of the whole does not (except in special cases) supervene on the operation of the individual parts. The grain of sand contributes to the mass of the sand pile. The kidney contributes to the capacity of creatures to maintain plasma osmolality. In each case, the property of the whole (the mass of the pile, the regulation of plasma osmolality) might differ even when the contribution of these singular parts remains the same. In *part-whole* levels, as opposed to levels of realization, the relationship is between parts and wholes, not between wholes and the corresponding organized collections of entities, properties, and activities. In this case, eliminativism about levels is a non-starter (whatever its metaphysical credentials); it is impossible to imagine neuroscience, biology generally, or indeed most special sciences without the idea that things have parts.

In applying the levels metaphor to this part-whole relation, one emphasizes the relations question over the placement question.¹¹ In levels of size, things are at a given level because they are similar in size. Levels, thus conceived, are *monolithic*: they reach across all of nature, embracing everything within a given size range. Oppenheim and Putnam's layered model of the world might be read as similarly monolithic. Wimsatt's tree diagram breaks with this monolithic view precisely because it emphasizes compositional relationships: different branching levels are required because different kinds of whole (cosmological objects, human societies) are composed in different ways. If one centers part-whole thinking in one's application of the levels metaphor, then the metaphor carries no particularly useful answer to the placement

question. One can offer only a necessary condition: two things are at the same level only if they are not related as part to whole. Given that most things are not related to one another as part to whole, the resulting idea of being "at" a part-whole level has little or no conceptual significance.

Levels of parts and wholes lack many of the features in Wimsatt's prototype of levels. Many of the features in that prototype appear to derive from the monolithic conception shown in Figure 3: causes, forces, and laws are most plausibly thought to cluster together on the assumption that size is relevant to which forces can act, that causal relations are expressions of forces acting, and that laws govern these interactions. But if one places the part-whole relation in the center of one's metaphor, then there is no reason to embrace an empirical association between being at a given level and having a proprietary set of causal interactions for that level. Levels of parts and wholes must also be correlated with size differences because parts can be no larger than the wholes they compose. But the size differences between levels of parts and wholes are an accidental consequence of the part-whole relationship itself, not part of defining what it is for things to be at different part-whole levels.

In the following subsections, my goal is to sketch some contours of the relevant notion of part and whole. I start with classical mereology only to make the point that this apparatus was not constructed with an eye to developing a descriptively adequate account of the levels described by science. The more we learn about the limits of these classical models for our present purposes, the more we place constraints on the relevant notion of levels that, as I and others have argued, is central to the explanatory structure of neuroscience and the special sciences generally: *levels of mechanisms* (Bechtel 1988; Bechtel & Richardson 1993; Craver 2001; Machamer et al. 2000).

6.1 Types and tokens of parts and wholes

The Gibson SG has two humbucker pickups. My Gibson SG has two humbucker pickups. Not all

¹¹ I am here using the terms "part" and "whole" in an intuitive and inclusive way. Much of the literature on the metaphysics of parthood is simply unrelated to the many senses of part and whole used in the theories of the special sciences. I am not thinking only of objects or sets, but also about events and temporal units. I sketch a more restrictive kind of part-whole relation below, but this remains an open question (see Sanford 1993).

Gibson SG's have two humbucker pickups. But that's the factory model, the central exemplar or prototype against which variations are evaluated as more or less "typical." Likewise, when we talk about *the* human brain or *the* frog kidney, we are talking about types: the human type of brain, the frog type of kidney. And we talk also about the parts these types of things typically have.

The monolithic, layer-cake image in the Oppenheim-Putnam hierarchy is a mereology of types: societies are formed of organisms, organisms of cells, cells of molecules, and so on. Wimsatt's tree also represents relationships between types. Crystals are made of crystalline solids, which are made of inorganic molecules, which are made of atoms. It is true, of course, that all cells are made of molecules and that all organs are made of cells. But is not true that all cells are at a higher level than molecules generally.

It seems natural and harmless enough to treat the part-whole relations among types as generalizations over relations between particular wholes and particular parts. When one says, "The human brain has two hemispheres and a corpus callosum," one asserts that having these parts is typical of human brains. One is warranted on the basis of such a claim (though not always correct) in asserting of a particular human brain that it has two hemispheres. That is, the relation among part and whole *types* derives from a more primitive, token relationship between particular parts and wholes. Type-level claims about part-whole relations assert that such part-whole relations regularly or typically hold in the individuals in the relevant reference class.

One of the many useful insights contained implicitly in the branching structure of Wimsatt's tree diagram (Figure 1) is that different types of wholes are made up of different types of parts and are naturally decomposed into different levels. Both the human brain and the frog kidney are organs, and both are made of cells, but the cells in each are different, and these cells are organized differently into higher-level components. If we look within the human brain, we find that different brain regions are composed of altogether different components

and exhibit more or less proprietary organization. Broadman mapped the brain by studying these differences in cytoarchitecture from one brain region to the next. The receptive field organization of the visual cortex has, to my knowledge, no companion in the organization of the amygdala or of the mammillary bodies. Different types of brain regions/systems are made up of different types of components: they have different part-whole levels.¹²

When we say that objects or processes of one type are parts of objects or processes of another type, we are asserting that *ceteris paribus* token objects of the one type are composed of token objects of the second type. Indeed, standard attempts to define the part-whole relationship with logical rigor are expressed in terms of relationships among token individuals (Varzi 2014).¹³

6.2 Mereology

Although Oppenheim and Putnam describe the layer-cake structure of science in terms of different types of objects, the mereological structure they use to support this picture is expressed in terms of tokens. Classical mereology provides a very general and content-neutral account of the part relation. It can be used equally well to express the sense in which

¹² Elephant hearts are parts of elephants, and puffin hearts are parts of puffins, and hearts are parts of organisms. Yet puffin hearts are not at a lower part-whole level than elephants, and elephant hearts are not at a lower part-whole level than puffins.

¹³ I shall not enter here into the difficult question of how token properties, processes, and objects are individuated in the biological sciences. Consider the spatial memory system in a rat. If we take this to be one thing over the life of the organism, then it will be composed of many different sets of parts over the course of its existence, like the ship of Theseus. If we take it instead to be the spatial memory system involved in learning the layout of one particular maze, which learning might be constituted by multiple trials and extended investigation, then again we will have a single higher-level system composed differently at different time slices. If we focus on a given instant in time, then no learning can occur; learning is a kind of change. For now, I simply note (along with Bechtel & Mundale 1999) that the appropriate mapping between such parts and wholes presupposes a criterion of individuation for the whole, and what counts or does not count as a part will be determined by whether it contributes to that whole, however specified. This is related to Marcus' (2006) thought that any token identity between levels presupposes a (non-dummy) sortal that fixes the individuation conditions of the relata in the same way. Carl Gillett (2013) has called attention to the need for different accounts of compositional relations for properties, processes, and objects. Here, I am glossing over these differences to keep the discussion simple.

the word apple has the letter “a” as a part, in which courage has judgment as a part, and in which apple pie has cinnamon as a part. Common axioms associated with classical mereology, including the mereology adopted by Oppenheim and Putnam (Rescher & Oppenheim 1955), include:

- 1) **Reflexivity:** Every object is a part of itself.
- 2) **Transitivity:** Every part of a part of an object is part of the object.
- 3) **Extensionality:** An object is completely determined by the set of its parts; i.e., for objects to be identical, it suffices that they have all and only the same parts.
- 4) **Summation:** Any pair of objects (x, y), is itself an object, z, which is their sum.

I list these constraints only to illustrate that classical mereology will take us only so far in the effort to characterize part-whole levels. This is a formal theory, abstracted entirely from the concerns of practicing scientists. This means that there are constraints in the scientific conception of parts and wholes that classical mereology need not honor. First, the levels that Churchland and Wimsatt describe are space-, structure-, and time-involving in ways that classical mereology need not be. The set of integers is part of the set of real numbers, and “Consider the Lobster” is part of Wallace’s corpus, but not in the same way that the glutamate receptor is part of the chemical synapse. The glutamate receptor takes up part of the space occupied by the synapse as a whole. Its opening is part of the extended process by which neurons communicate. None of this is expressed or intended to be expressed in the generic part-whole relation of classical mereology.¹⁴

¹⁴ Marr’s levels are not space-involving in this way. The algorithm is not located within the computation, it is not a substage of the computation, and it is not organized together with other parts in the service of the computation.

Second, although reflexivity is involved in certain theoretical applications of mereology, it has no application in thinking about such space-involving levels. It does no justice to the biological concept to assert that every hippocampus is a part of itself. If levels are defined as a relationship between a part and a whole, and everything is a part of itself, then everything is at both higher and lower levels than itself. The parts surely must be proper parts.¹⁵

Third, it has been noted that the transitivity axiom often fails to apply to functional parts of the sort that populate physiological and biological theories (Varzi 2014). Eric is part of the championship pool team, and Eric’s locks are part of Eric, but his locks are not part of the team. However, if one requires of a part (entity, property, or activity) that it must be *relevant* to the property or behavior of the whole, then one can retain the transitivity of this relation, at least in many contexts. *If we ask not about Eric, but about the motion of his arm as he wields his cue*, then his locks are clearly not relevant while the muscles gliding his arm steadily forward *are* relevant. So too are, in some sense, the molecules transmitted across the neuromuscular junction during his backstroke. When what counts as a part is filtered in each iteration by explanatory relevance relations (they are not mere spatial or temporal parts but working parts—parts that are involved in, contribute to, or make a difference to the property or activity of the whole), then the relationship is, in fact, transitive. The appearance of a failure of transitivity in functional systems trades, it seems, on failing to relativize the decomposition into parts by a highest-level target (explanandum) phenomenon; not all the spatiotemporal parts of an object or process are relevant to everything it does. It is only relative to a highest-level activity or property of the hierarchy under consideration that the lower level parts are visible as components—as working parts in the mechanism. *If we think not about the team, Eric, and his locks, but rather*

¹⁵ This is not a big departure from classical mereology; one could simply restrict one’s attention to proper parts. But the point underscores the fact that classical mereology was not developed with an eye toward understanding the sense in which pyramidal cells are parts of the hippocampus.

about the victory, the shot, and the muscular contraction, matters seem different.¹⁶ The contraction contributes to the shot, which contributes to the victory. The locks will not appear in this hierarchy, but the relevant parts will.¹⁷

The extensionality theorem holds that no two distinct objects share all and only their proper parts. A hippocampus and a bust of the Dalai Lama formed out of the same pyramidal cells, granule cells, etc. that compose the hippocampus are, according to classical mereology, one and the same object. But in biological systems, the organization of components is often (perhaps always) relevant to the properties and activities of the whole. Again, parts appear as parts only relative to a decomposition framed by reference to some highest-level property or activity. This is Kauffman's point, enshrined in Glennan's law: a mechanism is always a mechanism of a given phenomenon (Kauffman 1971; Glennan 1996). Thus Kauffman:

A view of what the system is doing sets the explanandum and also supplies criteria by which to decide whether or not a proposed portion of the system with some of its causal consequences is to count as a part and process of the system. Specifically, a proposed part will count as a part of the system if it, together with some of its causal consequences, will fit together with the other proposed parts and processes to cause the system to behave as described. (1971, p. 260)

The more general point is that there is an application of the levels metaphor that is not merely a part-whole relationship as specified in classical mereology, but one in which the parts are relevant (explanatorily and constitutively) to some property or activity of the whole.

One can make a similar point with respect to the summation axiom. This theorem allows

¹⁶ Clearly Eric's muscles are not part of the team, but this reflects only the fact that teams can have only certain kinds of part as members. If we look rather at an activity of the whole and ask what contributes to that, a different picture emerges.

¹⁷ Specific details about lower-level parts might be screened off in cases of multiple realization. Specific details about the parts might not be relevant. In that case, it would appear one must appeal to more abstract properties of the parts.

one to form arbitrarily many gerrymandered wholes out of disparate and unconnected parts with no spatial, temporal, causal, or functional unity. Lewis (1991) calls this "unrestricted composition": whenever there are some things, there is also a fusion of those things. The Yankees's starting rotation and the now disparate parts of my mother's old Chevy Vega together form a whole. This way of thinking about parts and wholes has little or no application in biology because such gerrymandered wholes don't do anything interesting (though such wholes will have aggregate properties of the sort discussed below). The whole in such gerrymandered collections typically doesn't play any explanatory role. And what goes for wholes goes for parts as well. According to this classical picture, it is perfectly legitimate to claim that my dog, Spike, has four parts: a front quarter, a hindquarter, and two midsections of approximately equal length. There is nothing to prevent this way of talking; but the parts revealed in this decomposition do not cut Spike at his joints. The biological decomposition finds joints at causal interfaces, and identifies parts with more or less isolable (nearly decomposable) subsystems (Simon 1962) that contribute to the behavior of the whole.

In short, many of the ideas central to classical mereology must be amended or restricted if they are to apply to the part-whole levels distinctive of biology, neuroscience, and the special sciences generally. At least some of the work can be done by restricting the part-whole relation by a *relevance condition* on biological part-whole: all the lower-level properties, activities, and organizational features of the parts are relevant to—contribute to—the property or activity of the whole.

6.3 Levels of organization: Aggregates and mechanisms

So let us focus on an application of the levels metaphor that is a part-whole relation and a (constitutive) relevance relation. I will not dwell here on the appropriate notion of relevance (see Craver 2007, see also Harinen 2014). For now, we can work with the idea that each part in

such a hierarchy (in addition to being spatially and/or temporally contained within the whole) plays a necessary but insufficient role within a collection of parts that are jointly sufficient (but possibly redundant) for a given explanandum phenomenon (Couch 2011). That is, relevant parts might usefully be thought of as constitutive insufficient, but necessary part of an unnecessary but sufficient condition (I) for the behavior of the mechanism as a whole (Mackie 1973).¹⁸

Again following Wimsatt (1997), we can distinguish two ways that spatiotemporal parts contribute to a property or activity of a whole: *aggregation* and *organization*. An aggregate property is literally a sum of the properties or activities of the parts. The current flowing through an ion channel, for example, is a sum of the charges carried by individual ions. The concentration of a volume of a fluid is a sum of the number of particles in that unit volume. Aggregative properties change linearly with the addition and removal of parts. And aggregative properties do not change as the parts are inter-substituted with one another. Some properties of the hippocampus, such as its mass, remain the same when the cellular constituents of the hippocampus are reorganized to represent His Holiness. Other properties of the hippocampus, such as its information processing capacities, are destroyed. For truly aggregative properties, spatial, temporal, and causal organization among the components is irrelevant.

¹⁸ Once we have made this adjustment, the *relata* in this relevance-merology are no longer objects but rather properties, activities/processes, or (as is more common in philosophical parlance) events. One does not explain the elephant; one explains why the elephant has large ears or how the elephant circulates its blood. One does not explain gasses; one explains their temperature and pressure. This point marks a significant departure from the mereological views of levels discussed above. Each of those applications of the levels metaphor focuses on objects or types of objects (societies, organisms, cells, and so on) as the *relata*, not on their properties, activities, and aspects of their organization. In many cases, the components picked out in a mechanistic decomposition fail to correspond to paradigmatic objects with clear spatial boundaries. The synapse, for instance, is composed of part of a presynaptic cell (the axon terminal), part of a postsynaptic cell (the dendrite or bouton), and a gap between them. What unifies these items into an object is their organized behavior: the pre-synaptic cell releases transmitters that traverse the cleft and act on the postsynaptic cell. Synapses are not cells or parts of cells, nor are they composed of cells. Rather, they are objects unified by their relevance to a given activity of the whole, such as chemical transmission.

Aggregates are rare. The masses of the individual grains in a sand pile do, in fact, depend on the spatial distribution of the other grains (if one takes relativity seriously). What is presumed to be a homogeneous concentration of a liquid can in fact have local concentration differences depending on how the ions are organized in different parts of the fluid. In the case of non-aggregates, the activity or property of the whole is not a simple sum of the properties of the individuals. Adding or removing parts (e.g., the human heart) can lead to dramatic changes in how the system (e.g., the body) works. And rearranging the parts and their activities in space and time can eliminate the explanandum phenomenon entirely (as would happen if one randomly swapped parts of the circulatory system for one another). This is all true because spatial, temporal, and causal organization are relevant to (make a difference to, partly constitute) the property of the whole.

I use the term “mechanism” permissively to describe non-aggregative compositional systems in which the parts interact and collectively realize the behavior or property of the whole. Mechanisms are by definition more than the sums of their parts: they have properties their parts do not have, and they engage in activities that their parts cannot accomplish on their own.

Most mechanisms with which I am familiar involve myriad part-whole relations, some of which are more aggregative in nature, and some of which are less so. Many things brains do, for example, involve the flux of ions across a membrane, which flux is closer to the aggregative than the mechanistic end of the organizational spectrum. Other things brains do (such as the developing grid cells in the entorhinal cortex) require precisely organized relations among the activities of cells in and around the entorhinal cortex. This organizational spectrum from aggregate to mechanism covers all the relations that go into levels of organization, the superordinate class.¹⁹

In levels of mechanisms, the *relata* are some activity or property of a mechanism as a

¹⁹ Levy (2013) calls attention to the fact that biological systems typically involve both aggregation and organization.

whole,²⁰ and the activities, properties, or organizational features of its components (its relevant parts and organization). Some component, X's φ -ing, is at a lower mechanistic level than S's ψ -ing if and only if X's φ -ing is a component in S's ψ -ing, that is, if and only if X's φ -ing is a relevant spatiotemporal part of S's ψ -ing. In levels of mechanisms (as opposed to aggregates) lower-level components are organized together to make up some behavior or property of the whole; in aggregates, the properties of the parts are summed.

Levels of mechanisms are represented in Figure 4. At the top is the activity of some mechanism as a whole (S's ψ -ing). S's ψ -ing is a behaving mechanism. Although one can speak of the mechanism and its activity separately (as when a mechanism stands inactive but ready to act), such separation in thought is artificial. Even the static mechanism is defined and sub-divided by reference to what it does. ψ is the topping-off activity of the mechanism for which all lower-level components are relevant. It can be idealized as an input-output relation, though this is an impoverished way of understanding phenomena (see Craver 2007). One level down are the activities and components, the X's φ -ing, which compose and are organized together to constitute S's ψ -ing.²¹ Below that is another iteration of levels: the ρ -ings of Ps organized such that one of the Xs φ s as it does. By organization, I mean that the parts have spatial (e.g., location, size, shape, and motion), temporal (e.g., order, rate, and duration), and active (e.g., feedback or other motifs of organization;

see Levy & Bechtel 2014) relations with one another by which they work together to do something they cannot do on their own. As noted above, the relationship between levels is a part-whole relationship filtered further by constitutive relevance (Craver 2005; Harinen forthcoming). In levels of mechanisms, parts are made into higher-level components by being organized spatially, temporally, and actively into something. In more aggregate compositional relationships, they are summed into higher levels.

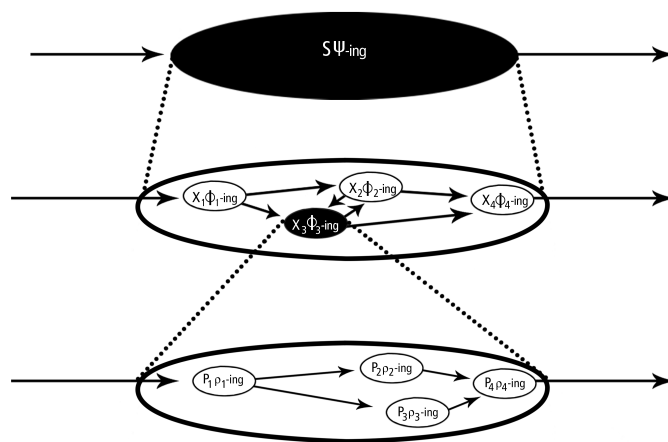


Figure 4: An abstract diagram of levels of mechanisms.

Contemporary theories of learning and memory provide a compelling example of levels of mechanisms (see Craver & Darden 2001; Craver 2002). The top level is a mechanism as a whole engaged in a spatial memory task, such as learning to run efficiently through a maze. One component in that mechanism, and so one level down in this description, is the hippocampus, a region of the brain thought to form a “map” of locations and orientations within the maze. The capacity of the hippocampus to acquire such an internal map of local spaces is thought to be explained, in part, by changes in synapses between pyramidal cells, specifically by a process known as Long-Term Potentiation (LTP). And it is now known that n-methyl d-aspartate (NMDA) receptors (n-methyl d-aspartate is a pharmacological agonist that binds these receptors preferentially), contribute to LTP. This story could continue downward, looking

²⁰ These might be understood as the obtaining of a property or the unfolding of a process over time. What counts as a static property often depends on one's temporal resolution.

²¹ I have not always chosen my language in a way that comports with common usage among metaphysicians, preferring to follow Salmon (1984). In this paper, I have tried to make it clear that I am interested in components. Components, as the name suggests, compose behavior of the higher-level mechanism when organized together. All of the component entities and activities organized together, it now seems appropriate to say, jointly constitute the behavior of the whole. That is, I am now using componency to talk about relationships between wholes and parts, and I am using “constitution” to talk about levels of micro-realization. I am not especially interested in the relationship between statues and lumps of clay. I am interested in how parts are organized and interact so that together they exhibit higher-level behaviors. I know of no metaphysician who has developed an adequate notion to express this, so perhaps I will be forgiven for appropriating these words for new uses.

into aspects of protein chemistry and the structural changes thought to underlie channel functioning.

6.3.1 Levels of mechanisms are local

Levels of mechanisms are of entirely local significance. The levels in our example are defined by reference to a topping-off point, spatial memory, contribution to which determines whether or not a spatiotemporal part of the system is in fact relevant—whether it is a component in the mechanism *for S's ψ -ing*. The hierarchy in Figure 4 and the levels of spatial memory as I have described them follow only a single (local) strand of embedding: from the behavior of the mechanism as a whole, to the behaviors of its components, on to the behaviors of *one* of these components, and so on.

Levels of mechanisms, like part-whole levels generally, are not monolithic divisions in the furniture of the world. Levels of mechanisms are defined only within a given part-whole hierarchy. There are different levels of mechanisms in the spatial memory system, in the circulatory system, in the osmoregulatory system, and in the visual system; the levels in each need not map onto one another. How many levels there are, and which levels are included, must be determined on a case-by-case basis by discovering which sorts of components are explanatorily relevant for a given phenomenon. Levels of mechanisms cannot be read off a menu of levels in advance.

If we apply the levels metaphor only locally, then it makes no sense to ask whether the hippocampus is at a higher or lower level than the nephra in the kidney. The nephra are not part of the hippocampus, and they are not relevant to the functioning of the hippocampus. Neither similarities of size nor similarities in kinds of part are definitive of levels of mechanisms. Rather, levels of mechanisms are defined relative to one another within a hierarchically organized mechanism.

The idea that levels of mechanisms retain some hint of the layer-cake model can sneak its way back into one's application of the metaphor

if one slides unknowingly between tokens and types of parts and wholes. Compare the following three sentences:

- a) This pyramidal cell is at a lower level of mechanisms than this hippocampus.
- b) Pyramidal cells are at a lower level of mechanisms than hippocampi.
- c) Cells are at a lower level of mechanisms than organs.

Statement (a) expresses a mechanistic notion of levels: a particular pyramidal cell is a component of a particular hippocampal mechanism. This statement is true if the cell is a component in a mechanism for a given activity in which the hippocampus is engaged. It might be, for example, that a given pyramidal cell is a component in some hippocampal mechanisms but not others; if so, it is at a lower level to some hippocampal activities and not others.

Wimsatt describes the compositional relationship between levels as a relationship between types. He writes: “Intuitively, one thing is at a higher level than something else if things of the first type are composed of things of the second type” (Wimsatt 1976, p. 215). This is a departure from the idea of levels of mechanisms and one that threatens to reinstate something like the Oppenheim and Putnam hierarchy. Pyramidal cells are found outside the hippocampus, and those pyramidal cells are not parts in hippocampal mechanisms; they are not at a lower level of mechanistic organization. Likewise, both the hippocampus and the kidney are composed of cells; organs tend to be composed of cells. But the cells in the hippocampus are not at a lower mechanistic level than kidneys because they do not contribute to kidney function. The slide from sentences such as (a) to sentences such as (b) and (c) is a slide back toward the layer-cake model. Of course, scientists typically deal with types. But as I suggest above, this is a generalization over a relationship between tokens. The correct generalization is that the cells that compose hippocampi are at a lower level than the hippocampi they compose.

This is significant for two reasons. First, it helps to show that many objections to thinking about neuroscience and other special sciences in terms of levels simply do not apply to this restricted application of the metaphor. If one thinks, with Wimsatt, me, and probably Oppenheim and Putnam, that the Oppenheim and Putnam layer cake is an overly simplistic representation of the diverse ontological structures one finds in the special sciences—that things like ocular dominance columns and synapses don't readily fit that picture—one can nonetheless retain the idea that mechanisms are susceptible to multiple nested decompositions. These are different applications of the levels metaphor. Secondly, and perhaps more importantly, the idea that levels are local significantly shifts the reductionist world-view for which Oppenheim and Putnam developed their ontology of levels. If one thinks of levels as levels of organization (as levels of mechanism and levels of aggregation), then it is inaccurate to think of reduction as involving relationships among theories developed to describe the items at a particular monolithic level. If reduction is simply a matter of explaining a higher-level phenomenon in terms of the organized activities of components, then reduction is still possible within a mechanistic world-picture, but it will be achieved not through grand reductions of overarching theories, but rather through piecemeal explanatory achievements for specific phenomena. Visions of the unity of science through interlevel reduction have to be revised not as grand unifications across the whole of science but rather as local explanatory successes. Such local explanations will, in fact, integrate findings from different sciences and bring different theoretical vocabularies into conversation with one another (see Craver 2005; Craver & Darden 2001), but it only deceptively resembles the layer cake that Oppenheim and Putnam sketched as a working hypothesis.²²

²² Nothing in this picture is meant to deny token identity between the behavior of a mechanism as a whole and the organized behavior of its parts. Because there are some conceptual difficulties that stand in the way of speaking meaningfully about token identities between levels, I have written with fewer commitments about constitution.

6.3.2 Placement is weak and derivative in levels of mechanisms

One consequence of the mechanistic application of the levels metaphor is that there is no unique answer to the question of when two items are at the same mechanistic level. Only a partial answer is available: X's ϕ -ing and S's ψ -ing are at the same level of mechanisms only if X's ϕ -ing and S's ψ -ing are components in the same mechanism, X's ϕ -ing is not a component in S's ψ -ing, and S's ψ -ing is not a component in X's ϕ -ing.²³ Unlike size levels or levels defined in terms of the types of objects found at a given level, levels of mechanisms are defined fundamentally by the relations question: by the componentency relationship between things at higher and lower levels. If two things are not related as part to whole, they are not at different levels, and so, if they are in the same mechanism, they are, in this very weak sense, at the same level. But this is just to say that sameness of level has no significance within this application of the metaphor.²⁴

If one thinks of levels of organization as levels of aggregates and levels of mechanisms, then spatial containment and size relations between levels follow as an accidental consequence of the componentency relation. The pyramidal cells are contained within the hippocampus, which is contained within the spatial memory system. The activities of these entities are also related as temporal part to whole: the binding of glutamate is a

²³ This has struck some readers as circular because it appears to state that X and S are at the same level if they are not at different levels. Appearances to the contrary, this is not circular. I have defined "same level" in terms of the notion of "different level" and the latter is defined in terms of componentency relations. The appearance of circularity, I believe, results from the fact that most people assume that the notion of "same level" must be primitive relative to the notion of "different level," and I have reversed that assumed order.

²⁴ Another way to see that levels of mechanisms do not answer the placement question is to recognize an apparent failure of transitivity. Suppose X1 and X2 are components in the same mechanism, that neither is a component in the other, and that the behavior of X2 can be decomposed into a set of interacting components, including P1. X1 would, according to this account, be "at the same level" as both X2 and P1 even though X2 and P1 are at different levels from one another. This problem, first raised by Lindley Darden (personal communication), is only a problem if one demands that there must be a unique answer to the placement question for an account of mechanistic levels. My argument against the notion of monolithic levels turns on the absence of any good principle for stretching the ideal of levels beyond its local context.

temporal component in the activity of the NMDA receptor. The objects at lower levels are smaller than (or at least no larger than) the whole, giving the hierarchy a derivative size ordering. Relations of size, rather than defining what it is for an item to be at a level (the placement question), are derivative from the more fundamental relationship between levels (the relations question): namely, the relationship between a mechanism and a component.

6.3.3 Emergence and levels of mechanisms

Mechanisms do things that their components taken individually cannot. This marks a sharp distinction between levels of mechanisms and levels of realization. Kim says this point is “obvious but important”:

This table has a mass of ten kilograms, and this property, that of having a mass of ten kilograms, represents a well-defined set of causal powers. But no micro-constituent of this table, none of its proper parts, has this property or the causal powers it represents. H₂O molecules have causal powers that no oxygen and hydrogen atoms have. A neural assembly consisting of many thousands of neurons will have properties whose causal powers go beyond the causal powers of the properties of its constituent neurons, or subassemblies, and human beings have causal powers that none of our individual organs have. Clearly then macroproperties can, and in general do, have their own causal powers, powers that go beyond the causal powers of their micro-constituents. (Kim 1998, p. 85)

Through aggregation or organization, wholes have causal powers that their parts individually do not have. An activity at a higher level of mechanistic organization is quite literally more than the sum of its parts. It is not an aggregate. The addition and removal of parts leads to nonlinear changes in the behavior of the mechanism as a whole. It matters how the parts are organized; it is in virtue of their organization that

they have properties that go beyond the properties of the individual parts (Wimsatt 1996). This feature of levels of mechanisms is so obvious, so prosaic, and so banal as to be hardly worth mentioning. No fancy complexity is required: two toothpicks stacked perpendicular to one another have the mechanistically emergent capacity to act as a lever or catapult; neither toothpick can do so on its own.

Of course, most mechanisms in biology are substantially more complicated than that. They have many more parts. Those parts interact with one another with bewildering complexity. Often they contain feedback relations that introduce nonlinear interactions into the operation of the mechanism itself. The mechanisms of LTP, for example, have yet to yield their secrets completely despite the dedicated attention of thousands of researchers over forty-odd years. A glance at any recent textbook is enough to convince one that LTP involves myriad intracellular reactions, protein synthesis, structural features of dendritic spines, changes to vesicular release, and retrograde transmission with nitric oxide. The mechanism involves so many parts and interactions that it would be useless, if not impossible, to represent them all in a visual diagram. Keeping track of how they all work together would require a very complicated computational model of some sort that has yet to be developed. As mechanisms get this complicated, we reach the limits of our ability to predict how the behavior of the whole will change as the parts change. Any change introduced to a part has so many ramifications that it is difficult or impossible for creatures like us to keep track of them all. This is an interesting fact about us and the limits of our cognitive and modeling prowess. But, ontologically, it is the same old banal fact about the importance of organization in mechanisms. We have added only that we have difficulty keeping track of it all.

Likewise, a common scientific complaint against reductionistic research programs in biology and neuroscience is that one can make only limited progress by studying the parts of mechanisms in isolation from one another.²⁵

²⁵ This idea of reduction is not the standard notion of theory reduction but something closer to what Eric Kandel means

We can study LTP in cells grown in a culture, and we can study hippocampal computations in a razor-thin hippocampal slice, and we can study spatial learning in highly contrived environmental settings such as a large pool filled with milky water (the Morris water maze). Such reductionist practices are absolutely essential to progress in the sciences. Nonetheless, one engaged in such practices must (and typically does) bear in mind that the behavior of the part when it is isolated for experimental purposes might be very different from the behavior the part exhibits when it is working in the context of a mechanism. Causal interactions with other parts of the mechanism and background conditions “in the wild” might lead to behaviors that would never be discovered in such simplified preparations. This is an extremely important point about reductionist research programs (Bechtel & Richardson 1993), and one might choose to describe this well-known difficulty with the language of emergence. But this is just to say that one cannot truly understand how a mechanism works until one understands how all its parts are organized together and working in the relevant conditions, and this we have already said repeatedly.

I emphasize the banality of these observations to stress that many of the things one wants to say about organization in biological systems can be said within the mechanistic application of the levels metaphor without introducing anything that is metaphysically novel or suspect. As the complexity of a mechanism increases, the epistemic challenges we face in discovering and modeling it increase as well, but this is of no significance for the ontic structures—the entities, activities, and organizational features that exist in the world.

Not so for *spooky emergence*. Spooky emergence is spooky precisely because it is committed to the existence of higher-level properties that have no explanation in terms of the parts, activities, and organizational features of the system in the relevant conditions. Levels of mech-

anisms are levels of ontic mechanistic explanation (Craver 2014): they are defined in terms of componency and constitutive explanatory relevance. If that explanatory relationship is severed, then the sense in which emergent properties are at a “higher level” must be altogether different than the compositional notion of levels in levels of mechanisms. If one imagines that atoms compose molecules, which are organized into cells, which are linked into networks from which mental properties spookily emerge, the first three steps are upward steps in a hierarchy of levels of mechanisms, but the last is not. The ability of organization to elicit novel causal powers (that is, nonaggregative behaviors and properties) is unmysterious both in scientific common sense and common sense proper (Van Gulick 1993; Kim 1998). Appeal to strong or spooky emergence, on the other hand, justifiably arouses suspicion. Indeed, it is unclear why properties that emerge in a spooky fashion should be thought of as higher-level at all. Perhaps the very idea of spooky emergence is incoherent.

6.3.4 Mechanistic levels are not causally related to one another

As with levels of realization, many common assumptions about the nature of causation would appear to make causal relations between mechanistic parts and the properties or behaviors of wholes suspect. Items at different levels of aggregation and at different levels of mechanisms are intimate with one another in much the same way that items at different levels of realization are intimate. Lewis is explicit. If C causes E:

C and E must be distinct events [if they are to be causally related]—and distinct not only in the sense of nonidentity but also in the sense of nonoverlap and non-implication. It won't do to say that my speaking this sentence causes my speaking this sentence; or that my speaking the whole of it causes my speaking the first half of it; or that my speaking causes my speaking it loudly, or vice versa. (Lewis 2000, p. 78)

by the term: choosing to study complex phenomena in extremely simple systems. We might call this experimental reductionism.

The relevant kind of intimacy for levels of mechanisms is overlap between token events or processes. The relationship between LTP and the opening of NMDA receptors during LTP induction is directly analogous to the relationship between speaking the whole of a sentence and speaking its first half. The induction of LTP is partly constituted by the opening of the NMDA receptor. The would-be cause in this top-down causal claim already contains the would-be effect within it. There is nothing additional to be produced in the effect; the occurrence of the effect includes the occurrence of the cause.

What about the bottom-up case? We might say that the spark plugs cause the engine to run, all the while acknowledging that the sparking of spark plugs is part of the operation of the engine. The naturalness of this locution is at least partly due to an ambiguity in the way we commonly describe the behavior of a mechanism as a whole. Sometimes we describe it as an activity or process that starts with the mechanism's setup conditions and ends with its termination conditions (Machamer et al. 2000). Thus we might describe Long-Term Potentiation as a *process*²⁶ or activity beginning with a rapid and repeated stimulus (called a tetanus) to the presynaptic neuron and ending with enhanced transmission across the synapse. Other times we describe the behavior of the mechanism as a whole, the phenomenon, as the *product* of that process (or one of its termination conditions). Thus we might say that the mechanism of Long-Term Potentiation produces a *potentiated synapse*. This way of speaking leads us to think in terms of the antecedent causes of potentiation: the tetanus is a distal cause, and the subsequent changes in the NMDA receptor are more proximal. If we think about the behavior of the mechanism in the second way, as a product, it is natural to think of the opening of the NMDA receptor as a cause of the synapses being potentiated (and indeed it is). But if we think about the behavior of the mechanism in the first way, as an input-output relation starting with the tetanus and ending with a potentiated synapse, then it is wrong to think of the tetanus or the opening of the NMDA receptor as a cause of *that*. The

NMDA receptor is a part of that causal process. These are two equally acceptable ways of describing the relationship between a mechanism and a phenomenon; they are easily translated into one another. However, if one is careless, these ways of speaking and writing invite equivocation of precisely the sort that we are struggling here to avoid.

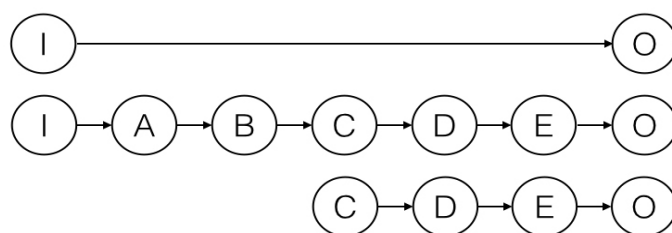


Figure 5: Why bottom-up causation is conceptually problematic.

Suppose we represent the input-output relationship constituting LTP as in the top of Figure 5, where I is a tetanus and O is a stable, potentiated synapse. Beneath this abstract I-O relation is a more detailed description of the intermediate stages in this mechanism: the tetanus excites the postsynaptic cell (A) which depolarizes it (B), causes NMDA receptors to open (C), and so on. (Nothing turns on the fact that I've idealized this mechanism as a single, step-wise causal chain.) Now, suppose we intervene to open the NMDA receptors directly and thereby potentiate the synapse (as shown in the third line). We might say that this intervention induced LTP; but when we say this, we mean that it produces the end product of the mechanism (it potentiates the synapse). We cannot coherently assert that it causes the process as depicted in the first two lines. This is for the simple reason that the process in the first two lines includes stages I, A, and B, and these are absent in the causal sequence represented in the third line. At most we can say the intervention caused the last half of the process. The NMDA receptor is not a cause of the process of LTP; it is a component of that process.²⁷

²⁷ This case is easiest to make for interventions that start the process mid-way. If an intervention, instead, were to augment C and thereby produce a more potentiated synapse than one would otherwise have had, then causal language would appear to be appropriate. The intervention changes, makes a difference to, the input-output relationship. These considerations generalize naturally to claims about types of

²⁶ Not in the Salmon (1984) sense, but in the colloquial sense of an unfolding sequence of states and activities.

6.4 Levels of mechanisms in relation to other kinds of levels

This application of the levels metaphor, according to which levels of organization are understood in terms of levels of aggregation and levels of mechanisms, thus offers a no-nonsense, ontological picture that comports well with the kinds of explanatory structure one finds in neuroscience and throughout the special sciences generally.

This view eschews the idea that levels are monolithic strata in the structure of the universe, with proprietary causal laws and forces (contra the view in Wimsatt, Oppenheim, and Putnam). Likewise, it allows that items at higher levels have causal powers that things at lower levels do not, in contrast to levels of realization. Single sciences and theories might investigate phenomena at many levels of organization, and an item located at one level in such a compositional hierarchy might be studied by many sciences and described in many theories. Things at different levels of organization (aggregation and mechanism) do not causally interact with one another, though we might find more complicated ways of describing how these items depend upon one another (see Craver & Bechtel 2007). As a result, if we think about the world in terms of levels of organization, we should not be tempted into thinking that things at higher levels control or dominate things at lower levels. Levels of organization are, in a sense, levels of explanation, given that explanations for different topping off phenomena will often decompose the system into altogether different parts within parts. It might be difficult to discern such levels in scientific practice, and the organization of components might be very complex, but nothing emerges from levels of mechanisms except in the banal sense that parts organized together do things that the parts alone cannot. Levels of organiz-

ation, in other words, seem to capture many of the intuitions that accompany the idea that the world is organized into levels but without many of the objectionable elements of other applications of the levels metaphor. The fact that the levels metaphor is often used carelessly and deployed in ways that violate common sense and metaphysical ideas about the structure of the world should not lead one to abandon the metaphor entirely. As we've seen, it can be given a relatively precise and metaphysically unobjectionable formulation that, in addition, fits the multilevel structures that the most advanced special sciences seem to be discovering.

7 Conclusion

Despite the ubiquity of levels talk in contemporary science and philosophy, very little has been done to clarify the notion. Here I defend a kind of descriptive pluralism about the levels metaphor: it is applied usefully in many contexts to describe different relata, different relations, and different senses in which items might be located at a given level. Because the levels metaphor is so ubiquitous and so promiscuously applied, some vigilance is required to keep the applications distinct from one another. I have discussed only a few applications: levels of science, theory, realization, size, mereology, aggregation, and mechanism. Even in these few key examples, we have found good reason to remain vigilant. The implications of the levels metaphor in one application only occasionally transfer when the metaphor is applied differently.

I have also suggested that levels of mechanisms (or, more generally, levels of organization) are especially important to the explanatory structure of neuroscience and the special sciences generally. If one thinks of levels in this way, one can easily see why interlevel causation should seem so problematic (indeed, it is problematic), one is free to jettison Oppenheim and Putnam's idea of monolithic levels of nature, and one can see room in the causal structure of the world for the existence and legitimacy of higher-level causes and explanations. Whether

mechanistic parts and wholes. Separately: When we describe this relation as a kind of production, levels show up as intermediate causes. Perhaps the temptation to speak of levels at all is lessened if one maintains that perspective. But this is not a change in what is being said so much as a change in how it is being said.

the idea of levels of mechanisms truly pays off in such useful ways remains to be seen. I merely hope to have preserved the metaphor, and its application to mechanisms, in the face of problems it inherits only through equivocation.

Acknowledgements

This essay is dedicated to William C. Wimsatt, whose sensitive and pioneering explorations of complexity, levels, and organization have shown how to tame the biopsychological thicket without sacrificing its wildness. Thanks to Anthony Dardis, Jens Harbecke, Donald Goodman-Wilson, Eric Hochstein, Eric Marcus, Lauren Olin, Gualtiero Piccinini, Anya Plutynski, Philip Robbins, Mark Povich, Felipe Romero, and Gary Williams for comments on earlier drafts.

References

- Baumgartner, M. (2010). Interventionism and epiphenomenalism. *Canadian Journal of Philosophy*, 40 (2), 359-384. [10.1353/cjp.2010.0015](https://doi.org/10.1353/cjp.2010.0015)
- (2013). Rendering interventionism and non-reductive physicalism compatible. *Dialectica*, 67 (1), 1-27. [10.1111/1746-8361.12008](https://doi.org/10.1111/1746-8361.12008)
- Bechtel, W. & Mundale, J. (1999). Multiple realizability revisited: Linking cognitive and neural states. *Philosophy of Science*, 66 (2), 175-207. [10.1086/392683](https://doi.org/10.1086/392683)
- Bechtel, W. & Richardson, R. C. (1993). *Discovering complexity: Decomposition and localization as strategies in scientific research*. Princeton, NJ: Princeton University Press.
- Bechtel, W. (1988). *Philosophy of science: An overview for cognitive science*. Hillsdale, NJ: Erlbaum.
- (2013). Addressing the vitalist's challenge to mechanistic science: Dynamic mechanistic explanation. In S. Normandin & C. T. Wolfe (Eds.) *Vitalism and the scientific image in post-enlightenment life science 1800-2010*. Dordrecht, NL: Springer.
- Bickle, J. (1998). *Psychoneural reduction: The new wave*. Cambridge, MA: MIT Press.
- Churchland, P. S. & Sejnowski, T. J. (1992). *The computational brain*. Cambridge, MA: MIT Press.
- Churchland, P. S. (1995). Can neurobiology teach us anything about consciousness? *Proceedings and Addresses of the American Philosophical Association*, 67 (4), 23-53. [10.2307/3130741](https://doi.org/10.2307/3130741)
- Couch, M. (2011). Mechanisms and constitutive relevance. *Synthese*, 182 (3), 119-145. [10.1007/s11229-011-9882-z](https://doi.org/10.1007/s11229-011-9882-z)
- Craik, F. I. M. & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning & Verbal Behavior*, 11 (6), 671-684. [10.1016/S0022-5371\(72\)80001-X](https://doi.org/10.1016/S0022-5371(72)80001-X)
- Craver, C. F. & Bechtel, W. (2007). Top-down causation without top-down causes. *Biology and Philosophy*, 22 (4), 547-563. [10.1007/s10539](https://doi.org/10.1007/s10539)
- Craver, C. F. & Darden, L. (2001). Discovering mechanisms in neurobiology: The case of spatial memory. In P. K. Machamer, R. Grush & P. McLaughlin (Eds.) *Theory and method in neuroscience* (pp. 112-137). Pittsburgh, PA: University of Pittsburgh Press.
- Craver, C. F. (2001). Role functions, mechanisms and hierarchy. *Philosophy of Science*, 68 (1), 31-55.
- (2003). The making of a memory mechanism. *Journal of the History of Biology*, 36 (1), 153-195.

- [10.1023/A:1022596107834](#)
- (2005). Beyond reduction: Mechanisms, multifield integration, and the unity of science. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 36 (2), 373-396. [10.1016/j.shpsc.2005.03.008](#)
- (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford, UK: Clarendon Press.
- Darden, L. & Maul, N. (1977). Interfield theories. *Philosophy of Science*, 44 (1), 43-64.
- Darden, L. (1991). *Theory change in science: Strategies from mendelian genetics*. Oxford, UK: Oxford University Press.
- Fehr, C. (2004). Feminism and science: Mechanism without reductionism. *National Women's Studies Association Journal*, 16 (1), 136-156. [10.1353/nwsa.2004.0032](#)
- Floridi, L. (2008). The method of levels of abstraction. *Minds and Machines*, 18 (3), 303-329. [10.1007/s11023](#)
- Foster-Wallace, D. (2004). Consider the lobster. *Gourmet Magazine*, 64 (8)
- Gillett, C. (2002). The dimensions of realization: A critique of the standard view. *Analysis*, 62 (276), 316-323. [10.1111/1467-8284.00377](#)
- (2013). Constitution, and multiple constitution, in the sciences: Using the neuron to construct a starting framework. *Minds and Machines*, 23 (3), 209-337. [10.1007/s11023-013-9311-9](#)
- Glenan, S. (1996). Mechanisms and the nature of causation. *Erkenntnis*, 44 (1), 49-71. [10.1007/BF00172853](#)
- Hafting, T., Fyhn, M., Molden, S., Moser, M. & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436 (7052), 801-806. [10.1038/nature03721](#)
- Harinen, T. (forthcoming). *Causal and constitutive explanation*.
- Haug, M. C. (2010). *Realization, determination, and mechanisms*. *Philosophical Studies* 150 (3), 313-330.
- Hempel, C. G. (1965). *Aspects of scientific explanation*. Princeton, NJ: Princeton University Press.
- Hitchcock, C. (2003). Of Humean bondage. *The British Journal for the Philosophy of Science*, 54 (1), 1-25. [10.1093/bjps/54.1.1](#)
- Kauffman, S. A. (1971). Articulation of parts explanation in biology and the rational search for them. In R. C. Buck & R. S. Cohen (Eds.) *PSA 1970* (pp. 257-272). Dordrecht, NL: Reidel.
- Kim, J. (1998). *Mind in a physical world*. Cambridge, MA: MIT Press.
- (2000). Making sense of downward causation. In P. B. Andersen, C. Emmeche, N. O. Finnemann & P. Voetmann Christiansen (Eds.) *Downward causation. Minds, bodies and matter* (pp. 305-321). Aarhus, DK: Aarhus University Press.
- Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In P. Kitcher & W. Salmon (Eds.) *Scientific explanation* (pp. 410-505). Minneapolis, MN: University of Minnesota Press.
- Lewis, D. (1991). *Parts of classes*. Oxford, UK: Blackwell.
- (2000). Causation as influence. *Journal of Philosophy*, 97 (4), 182-197.
- Machamer, P. K. & Sullivan, J. (2001). Leveling reduction. *University of Pittsburgh Philosophy of Science Archive*
- Machamer, P. K., Darden, L. & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, 57 (1), 1-25. [10.1017/CBO9780511498442.003](#)
- Mackie, J. L. (1980). *The cement of the universe: A study of causation*. London, UK: Clarendon Library of Logic and Philosophy.
- Marcus, E. A. (2006). Events, sortals, and the mind-body problem. *Synthese*, 150 (1), 99-129. [10.1007/s11229-004-6258-7](#)
- Marr, D. (1982). *Vision*. San Francisco, CA: Freeman Press.
- Melnyk, A. (2010). Comments on Sydney Shoemaker's "Physical realization". *Philosophical Studies*, 148 (1), 113-123. [10.1007/s11098-010-9500-9](#)
- Moser, E. I., Kropff, E. & Moser, M. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annual Review of Neuroscience*, 31, 69-89. [10.1146/annurev.neuro.31.061307.090723](#)
- Oppenheim, P. & Putnam, H. (1958). Unity of science as a working hypothesis. In H. Feigl, M. Scriven & G. Maxwell (Eds.) *Concepts, theories, and the mind-body problem, Minnesota studies in the philosophy of science II* (pp. 3-36). Minneapolis, MN: University of Minnesota Press.
- Polger, T. (2004). *Natural minds*. Cambridge, MA: MIT Press.
- Rescher, N. & Oppenheimer, P. (1955). Logical analysis of gestalt concepts. *British Journal for the Philosophy of Science*, 6 (22), 89-106. [10.1093/bjps/VL22.89](#)
- Romero, F. (forthcoming). Why there isn't interlevel causation in mechanisms. *Synthese*.
- Salmon, W. (1984). *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton University Press.

- Samsonovich, A. & McNaughton, B. (1997). Path integration and cognitive mapping in a continuous attractor neural network model. *Journal of Neuroscience*, 17 (15), 5900-5920.
- Sanford, D. H. (1993). The problem of the many, many composition questions, and naive mereology. *Noûs*, 27 (2), 219-228. [10.2307/2215757](https://doi.org/10.2307/2215757)
- Schaffner, K. F. (1993). *Discovery and explanation in biology and medicine*. Chicago, IL: University of Chicago Press.
- Shagrir, O. & Bechtel, W. (forthcoming). Marr's computational-level theories and delineating phenomena. In D. Kaplan (Ed.) *Integrating psychology and neuroscience: Prospects and problems*. Oxford, UK: Oxford University Press.
- Shagrir, O. (2010). Computation, San Diego style. *Philosophy of Science*, 77 (5), 862-874. [10.1086/656553](https://doi.org/10.1086/656553)
- Shepherd, G. (1994). *Neurobiology*. London, UK: Oxford University Press.
- Simon, H. (1969). *The sciences of the artificial*. Cambridge, MA: MIT Press.
- Thalos, M. (2013). *Without hierarchy: The scale freedom of the universe*. Oxford, UK: Oxford University Press.
- Van Gulick, R. (1993). Who's in charge here? And who's doing all the work? In J. Heil & A. Mele (Eds.) *Mental causation* (pp. 233-256). Oxford, UK: Oxford University Press.
- Varzi, A. (2014). Mereology. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. <http://plato.stanford.edu/archives/spr2014/entries/mereology/>.
- Wilson, R. A. & Craver, C. F. (2006). Realization. In P. Thagard (Ed.) *Elsevier handbook of the philosophy of psychology and cognitive science* (pp. 81-104). Dordrecht, NL: Elsevier.
- Wimsatt, W. (1974). Complexity and organization. In K. F. Schaffner & R. S. Cohen (Eds.) *PSA 1972* (pp. 67-86). Dordrecht, NL: Reidel.
- (1976). Reductionism, levels of organization, and the mind-body problem. In G. Globus, I. Savodnik & G. Maxwelll (Eds.) *Consciousness and the brain* (pp. 199-267). New York, NY: Plenum Press.
- (1997). Aggregativity: Reductive heuristics for finding emergence. *Philosophy of Science*, 64 (4), 372-384. [10.1086/392615](https://doi.org/10.1086/392615)

Mechanistic Emergence: Different Properties, Different Levels, Same Thing!

A Commentary on Carl F. Craver

Denis C. Martin

In this commentary I will briefly sketch the notion of “levels of mechanisms” as presented by Carl Craver and propose that we extend it to a more general notion of “level” that ensures wider applicability. The account of levels I develop is essentially based on an account of “properties”, claiming dependence of instantiation on a certain epistemic context. The main goal is then to reconcile Craver’s notion of “mechanisms” with “emergence” resulting in a contemporary account of “mechanistic emergence” implementing the developed concept of a level. Such an account could provide explanatory potential for and elucidate on seemingly mysterious higher-level properties and their ontology.

Keywords

Causation | Descriptive pluralism | Dispositional essentialism | Dualism | Emergence | Epistemic context | Level | Mechanisms | Mechanistic emergence | Mechanistic explanation | New essentialism | Novelty | Ontological novelty | Part-whole relationship | Properties | Property instantiation | Realization | Unexplainability | Unpredictability

Commentator

Denis C. Martin

denis.martin@hu-berlin.de

Humboldt Universität zu Berlin
Berlin, Germany

Target Author

Carl F. Craver

ccraver@artsci.wustl.edu

Washington University
St. Louis, MO, U.S.A.

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

Are mind and brain on the same level? Mental properties and biological properties are so different that some kind of dualism is still an attractive position for many people. Intuitively, mental phenomena are often assumed to be on some kind of higher level than physical phenomena. For example, in order to accurately describe what it means to have compassion for another living being, most people would probably

agree with the popular expression that this simply *cannot amount to nothing but* the description of the underlying neurophysiological activity or behaviour *related* to that compassion—that presenting the neurophysiological activity alone does not fully capture all properties of being in a state of compassion. Instead, especially in everyday life, we might rather refer to the phenomenological properties of compassion,

the properties we draw on to identify that we are in a state of compassion at a given time. These properties seem to have a special value for us. In a way, they seem to be much richer than those of “cold” science. But what, exactly, does it mean to say that mental and physical phenomena are not on the same level? If I were to ask what compassion is, most people would probably agree that it is somehow realized by their body, just as an elaboration of this fact does not suffice for a complete description of compassion, implying that there must be something *more* than that, on a *higher* level. This, at least for a philosopher, inevitably leads to the question of what those levels actually *are*. What does “level” refer to? To what extent do levels *exist* in the world at all? These questions become even more pressing when we make ourselves aware of the extent to which the sciences use the concept of “level”. Whole disciplines, such as psychology and neuroscience, are distinguished as operating on different levels with different theories aiming at specific target phenomena. Levels also play a role within disciplines. In neuroscience, for example, it is quite common to distinguish between lower-level brain functions as realized in the brain stem or primary sensory areas as opposed to higher-level functions like decision-making or emotion regulation that are attributed primarily to the frontal lobe. Likewise, the distinction of processes and functions as operating “bottom-up” or “top-down” is quite prevalent.

There is a general strategy in science that has proven to be effective for explaining a certain phenomenon: decomposition. The reason for that is as follows: to fully explain a phenomenon, it does not suffice for us to be able to elaborately describe it or list certain correlations with singular components or other phenomena. Rather, we need to know in detail how the phenomenon comes into existence, based on how exactly it is realized: which components underlying the phenomenon are doing what, where, when, and how in order to make the phenomenon emerge. These requirements are captured excellently by Carl Craver’s (2007; Craver & Darden 2013, p. 15) famous definition of a mechanistic explanation:

mechadef/mechanistic explanation =_{Df} [m]echanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions. (Craver & Darden 2013, p. 15)

But in what sense are the mechanistic components of a phenomenon on a lower level than the whole phenomenon? This is the question Craver answers in his article “Levels”, in this collection.

In what follows, I shall first point out I find most important about Craver’s account of levels of mechanisms and where I see some difficulties in his account. I shall then propose an alternative way of defining levels by emphasizing the notion of “properties”. The idea here is that levels are a direct result of property instantiations and thereby constitute “property-dependent epistemic dimensions”. By focusing on properties in general, and not only on properties of mechanisms, I hope to show that an account of levels does not have to be as restricted as Craver proposes. I shall also argue that levels of mechanisms and levels of emergence do not have to be conceived as necessarily distinct, but can rather be combined quite well into a productive account of mechanistic emergence. Expanding Craver’s account of levels this way provides not only a notion of levels with wider applicability but also builds on his account of mechanisms as operating on different levels, which instils explanatory potential into a contemporary account of emergence. This still secures the application of “levels” in science, but at the same time makes transparent how the epistemic contexts of science are property- and level-generative. The ultimate goal of this approach is, of course, to elucidate how one and the same material system may show significantly different properties to an extent that elicits serious confusions about matters of identity.

2 Levels of mechanisms

First of all, the approach Craver takes for defining levels is notable in many ways. What I find especially important, however, is that he develops his definition in an interdisciplinary frame-

work, paying close attention to compatibility with or even application in neuroscience. So what are levels as used in such a scientific context? In his article of this volume, Craver extends his original definition of mechanisms a little to accommodate for the existence of lower-level mechanisms that take part in the realization of higher-level mechanisms.

I use the term ‘mechanism’ permissively to describe non-aggregative compositional systems in which the parts interact and collectively realize the behavior or property of the whole. Mechanisms are by definition more than the sums of their parts: they have properties their parts do not have, and they engage in activities that their parts cannot accomplish on their own. (Craver [this collection](#), p. 16)

Mechanisms as construed here are entities and activities organized in non-aggregative compositional systems, such that they are productive of regular changes from start or set-up to finish or termination conditions, and the properties of the whole mechanism are produced collectively through the interaction of its component mechanisms. This establishes the basis for Craver’s introduction of levels of mechanisms.

Craver’s three defining questions—namely about the relata, the relations, and their placement—constitute a valuable contribution to the conceptual clarification of the term “level” helping to capture the criteria for the correct usage of the term. This already hints at the descriptive pluralism Craver promotes, meaning that there is a set of equally correct ways to use the term depending on the respective answers to these three key questions. In the case of mechanisms, levels are best individuated, according to Craver, in terms of a part–whole relationship between the property “ ψ -ing” of S, given that S is a mechanism as a whole, and the property “ ϕ -ing” of X, given that X is a mechanism component that is a constitutively relevant spatiotemporal part of S.

Summing up Craver’s position on levels and levels of mechanisms as I understand it, they:

- are *metaphors* with multiple distinct conditions of correct usage dependent on the relata, relations, and their placement and
- refer to *part–whole relationships*.
- Levels of mechanisms have *properties* of mechanisms and properties of their parts *as their relata* (as opposed to levels of size, which have objects as their relata),
- are always *non-aggregative* (though aggregative levels do also exist),
- are *not monolithic*, but constitute a *local organizational part–whole hierarchy*, while
- the part–whole relationship must satisfy the *constitutive relevance condition*;¹ **while**
- there is *no causal relation* between them,
- they bear *explanatory potential*,
- and, finally, the *placement* of entities on levels of mechanisms *is weak* in the sense that for all entities that are not related as part and whole it can be said that they are on the same level.

What might the difficulties with this account of levels be? As a minor point, first, there might be some implications of using the concept “metaphor” in connection with levels. By definition, a “metaphor” identifies two things—a primary and secondary subject—with one another, such that one of the two can be captured in description more powerfully (Hills 2010). What could the primary and secondary subject in a level-metaphor be? The primary subject would probably be a level in the sense of a level of mechanisms, the secondary subject could maybe be a plane, a stage, or a degree. But how would that help elucidate what the primary subject levels actually are? As far as I can see, using “level” as a metaphor would more effectively *describe* what a level is, but not actually *define* it and, thereby, simply capture what our intuition about levels is in the first place—namely that it is somewhat analogous to a level in the secondary subject sense. Also, it seems that conceiving of “levels” as a metaphor would

¹ “Constitutive relevance”: “[...] all the lower-level properties, activities, and organizational features of the parts are relevant to—contribute to—the property or activity of the whole” (Craver [this collection](#), p. 15).

already somewhat negatively answer the question of what levels actually are, as solely existent as a figure of speech—an analogy that could be eliminated without any ramifications. Craver’s descriptive pluralist approach is formulated specifically to counteract elimination of levels and thereby to sustain their application in science. However, descriptive pluralism obviously does still act on the assumption that levels are metaphors, and only describes conditions that fit their application better than others. Therefore, this approach does not seem to be particularly helpful for our intended and certainly desirable goal.

Another issue I would like to raise is that in Craver’s account of levels as presented here, the key defining feature of levels seems to be a relation condition² between certain entities. This already becomes apparent with the three questions mentioned above, which aim to help us adequately describe specific instances of levels. Levels of mechanisms in particular are specified as a part–whole relationship between properties of mechanisms that are located on different levels. But does this really capture what *levels* actually are? Or does it rather render the relation condition *between* levels more precise, instead? While helpful to set the criteria for conditions under which the term “level” is correctly employed, and highlighting distinguishing features of different levels, saying that levels are essentially relations between sets of entities is at best an indirect or descriptive definition of levels and does not seem sufficient for a complete definition. It leaves open how levels come into existence, what their ontological status is, and why we posit certain entities on the same level in the first place, that is, what the commonalities of entities are that lead them to be on the same level.

From the key aspect of a part–whole relationship in Craver’s account stems the notion that levels are local and non-monolithic. This means that only those entities that are involved in such part–whole relationships can intelligibly be said to be on different levels. What are our theoretical options for conceptually covering all

other entities? Since they don’t fulfil the part–whole criterion they cannot be on distinct levels and, therefore, in a sense they are all on the same level. However, according to the definition of levels at hand, to be on any level means that there are other levels, which are distinguished from the first level by the part–whole relationship of the entities involved. Since the entities under consideration do not exhibit this kind of relationship, they are on no level at all. In general, this seems like a reasonable option. But let’s consider the case that entities that are not in such a part–whole relationship do, nevertheless, share some features—for example a hedge and a fence both one meter high. In accordance with levels not being monolithic and the previous considerations, these would not be on the same level, but on no level at all. What exactly is wrong, however, with saying that two entities sharing the feature of being one meter high but which are not related as part and whole are on the same level? One could, of course, simply invoke the account of levels of mechanisms and argue that it is designed such as not to warrant such a level. But does this limitation really procure us a better understanding of “level”, or could it rather be too restrictive for that purpose? Its consequence, at any rate, is a very strong focus on the vertical dimension, namely the relation between levels, whereas the horizontal dimension, that is, entities related qua being on the same level, is somewhat neglected. So, let us ask, what are the criteria for two entities being on the same level? It is exactly this relation between entities on the *same* level that the concept is primarily supposed to capture, and yet which seems to be underspecified by the definition provided. And how similar do two part–whole relationship units have to be in order for it to be correct to say that their respective wholes and parts are on the same level? Or does it even follow that two things that are not part of one and the same part–whole relationship cannot even be on the same level at all?

As a third point, finally, there remains the issue of the extent to which levels of mechanisms are similar or distinct from levels of realization and emergence, respectively. All three kinds of levels share that an application of the

² “Relation condition”: “the componency relationship between things at higher and lower levels” (Craver [this collection](#), p. 19).

concept of inter-level causation is not feasible in their case, since they do not fit classical assumptions about causation such as non-synchronicity of cause and effect—a very substantial point Craver emphasizes in the target article. But what are the differences between these kinds of levels?

According to Craver, levels of mechanisms and levels of realization seem to differ in that the former exhibit a relationship between wholes and parts, whereas the latter exhibit a relationship between wholes and sets of realizers. But this distinction seems rather frail. How are the parts involved in levels of mechanisms different from the set of realizers involved in levels of realization, such as to warrant this distinction? At least in levels of mechanisms, as Craver envisages them, the parts are several mechanisms that together form the whole, which is comprised of all the particular “part-mechanisms”. If mechanisms in general realize certain phenomena, this suggests that all “part-mechanisms” are also realizers in the same way. Now, if the “part-mechanisms” on the “part-levels” are all realizers, it is reasonable to say that the “whole-mechanism” on the “whole-level” is realized by the organized coaction of its parts and that the “part-levels” are the realizers of the “whole-level”. Thus, the distinction between levels of mechanisms and levels of realizers conflates.

The difference between levels of mechanisms and levels of emergence, on the other hand, is based on the unpredictability, unexplainability, and metaphysical novelty of higher-level properties, as opposed to lower-level ones. Craver’s point here is that levels of mechanisms, while they can be unpredictable, do not have to be so necessarily, that they are always explanatory, and the novelty of higher-level properties is a trivial fact. But why think that the opposite must hold in the case of emergence? Of course, “spooky emergence”, “[...] the existence of higher-level properties that have no explanation in terms of the parts, activities, and organizational features of the system in the relevant conditions” (Craver [this collection](#), p. 21), is spooky by definition—that much is clear. Also, admittedly, the way emergence was construed

historically by the British Emergentists perfectly fits this view and deliberately opposes mechanisms as it can be found, for example, as per Broad (1925). However, why should we prematurely accept this view of emergence as given and eliminate any possibility of further development towards a notion of emergence that is perfectly commensurable with modern science? In fact, I think the formidable way in which Craver develops his account of levels of mechanisms is perfectly suited to facilitate development in this direction. So what could a definition of emergence be, and how can it be united with mechanisms? The following definition of emergence by Evan Thompson (2007) already seems compatible with Craver’s framing of the way properties of higher-level mechanisms are constituted by properties of lower-level ones.

A network, N , of interrelated components exhibits an emergent process, E , with emergent properties, P , if and only if:

- (1) E is a global process that instantiates P , and arises from the coupling of N ’s components and the nonlinear dynamics, D , of their local interactions.
- (2) E and P have a global-to-local (“downward”) determinative influence on the dynamics D of the components of N .

And possibly:

- (3) E and P are not exhaustively determined by the intrinsic properties of the components of N , that is, they exhibit “relational holism.” (p. 418)

This definition is compatible with Craver’s characterization of levels of mechanisms in the following respects: properties of higher-level mechanisms, global emergent properties, are realized by properties of lower level mechanisms; and there is a part-whole relationship between those relata, as well as a non-causal influence between the levels. What Thompson’s definition additionally contributes is a point about predictability. For many phenomena in

nature, the interactions of lower-level components are so complex that they can only be described by non-linear dynamics. A precise predictability of the higher-level phenomena might not always be possible at present due to there being too many factors involved in the underlying processes—it simply exceeds current computational tractability. Craver acknowledges this point—so it can also be said to be consistent with his account—but worries that this might have ontological ramifications: “[i]f that explanatory relationship is severed, then the sense in which emergent properties are at a ‘higher-level’ must be altogether different than the compositional notion of levels in levels of mechanisms” (Craver [this collection](#), p. 21). More precisely, he suspects that in emergence ontological novelty arises through the epistemological limitations just mentioned; otherwise ontological novelty would simply be a banal fact already expressed by his account.

There are, however, several problems with this view: first, one can make a distinction between the epistemological (e.g., “predictability”) and ontological (e.g., “novelty”) dimension of emergence (O’Conner & Wong 2009). There is, in principle, no reason to assume that the ontological dimension is dependent on the epistemological; rather, they seem to be fully dissociable.

Second, his criticism backfires with regard to the banality of the properties that higher-level mechanisms exhibit in his own account of levels. For there to be “higher” levels of mechanisms, these mechanisms must show *new properties*, that is, in order for them to qualifiedly be on that level. Hence, his account cannot go without ontological novelty of some kind. To now say that this ontological novelty would be only a banal fact undermines his very own striving for mechanistic explanation, which certainly is not banal. In fact, it is still interesting how “higher-level” properties come into existence, what it means to say that they are *new*, and how the concept “level” might be connected to this. A successful reconciliation of mechanisms and emergence in the form of mechanistic emergence could provide a solution to this problem.

Third, the dissociability of the epistemological and ontological dimension of emergence does not contradict the possibility of their mere coexistence. Once we dismiss the idea that ontological novelty follows from epistemological intractability, overcoming the restraints of historical accounts of emergence, the fact that the coming into existence of new properties on a higher level is not tractable at the moment does not mean that it is not so *in principle*. The reason why we call this coming-into-existence “emergence”, as might be conceived by a revised account, is not based on the fact that it *is* epistemically intractable *in principle*, but rather that it shows novel properties on a higher level that *appear* to be epistemically intractable *in principle*, while they might at some point be very well explained in a mechanistic framework combined with a proper theory of property instantiation.

Thompson’s definition leaves room for local components to be part of a mechanism. A mechanistic explanation of the emergent phenomenon, it seems, would not be incompatible with an account of emergence, but rather contribute to its explanation by elaborating on how the organization of the parts is essential for emergent properties to arise. As for the ontological novelty of higher-level phenomena, this is certainly a crucial point in emergence: there are new properties coming into existence on the higher levels that are somehow realized by processes of components on the lower level, which in isolation do not show the same properties as the whole. For this to happen, however, contrary to what Thompson’s definition implies, the underlying interactions of the components or the emergent properties themselves must not necessarily be unpredictable in principle. But I anticipate an objection: this form of emergence would again only be very weak or banal, but not ontologically new. As already mentioned above, unpredictability does not have a bearing on ontological novelty and is therefore not crucial for emergence. What levels of mechanisms and emergent levels share is that on higher levels there are *new properties*, which means that there is a notable ontological difference. How extensive such an ontological difference

must be in order not to be banal remains a matter of debate. Still, in the case of mechanisms, as well as in the case of emergence, it is very likely that there is a significant phenomenon making up the higher level—otherwise it probably would not be of such interest for enquiry as it clearly is.

As a result of these thoughts, in what follows I will try to reconcile levels of mechanisms and levels of emergence as two interconnected forms of realization. This alone, of course, does not solve the problem of what the levels involved actually *are*. So in fact there are two problems to be solved for an account of mechanistic emergence: (a.) what it means exactly for a higher-level phenomenon to exhibit *new* properties, and (b.) what exactly constitutes a *level*. As a route to a possible solution, in the next section I will sketch a definition of properties that can be implemented in a definition of mechanistic emergence and that at the same time provides a positive account of levels.

3 Level-carving properties in mechanistic emergence

In this section, I propose an alternative account of levels that is fully compatible with the mechanistic framework and the way in which levels of mechanisms are construed by Craver, but which at the same time has a wider scope of application. Since this account will rely on properties as the crucial defining criterion, I shall first sketch a working definition of the concept of “properties”. In a second step, this definition of “property” will conceptually ground the alternative account of levels. Finally, I will implement both the definition of properties and that of levels into a formulation of mechanistic emergence.

What might be a working definition for the term “properties”? Inspired by [Brian Ellis’ \(2002\)](#) “new essentialism”³ and [Alexander Bird’s \(2007\)](#) “dispositional essentialism”,⁴ I

propose the following view: what exist in the world are entities with individual dispositional profiles. An example that [Ellis \(2002\)](#) gives is the dispositions of particles to attract or repel each other. These essential dispositions, individuating the particles as that which they are, make it possible for us to formulate laws of nature. Many of these dispositions are the result of structural and organizational combinations of matter with different dispositions, e.g., ions have the disposition to form ionic crystals, which by means of the resulting structural characteristics in turn have other specific physico-chemical dispositions. Those essential dispositions alone, however, are not properties yet. That is because dispositions exist outside of an epistemic context. The property of being one meter high, for example, is dependent on the disposition of an object to exactly fit the measurement revealing it to be the height of one meter. Without the measurement, however, the property is not instantiated—only the disposition of its instantiation exists inherently in the structure of the object. Thus, according to my theory, properties are instantiated through the interaction of the essential dispositions of matter and an epistemic system. Of course, now you will ask what this epistemic system could possibly be. Admittedly, this aspect of the definition is in a particularly embryonic stage and requires further research. As a general characterization, epistemic systems are structured such that they feature sensors or gauges that capture specific dispositions of entities and provide characteristic values as an output. Human and non-human animals, as well as physical devices of measurement, are epistemic systems in this sense. Let us note the following as a working definition of “epistemic system”:

Epistemic system =_{DF} (ES) Epistemic systems are organized (a.) such as they feature sensors or gauges that pick up a specific disposition exhibited by an entity and (b.) such as there is a transformation of that signal into a particular value characteristic of the system’s organization.

³ “[...] things must behave in the sorts of ways they do not because the laws of nature require them to, but rather because this is how they are intrinsically disposed to behave” ([Ellis 2002](#), pp. 3–4).

⁴ “[...] the claim that fundamental natural properties are essentially dispositional. [...] *x* is disposed to manifest *M* in response to stimulus *S* iff were *x* to undergo *S* *x* would yield manifestation *M*” ([Bird 2007](#), p. 24).

With this definition at hand we are able to formulate a working definition of properties:

Properties =_{Df} (P) Properties are instantiated through epistemic processes, which are constituted by interactions between epistemic systems and complementary dispositional profiles of entities.

Let us now turn to how levels depend on properties. The idea here is that levels are established by the epistemic systems in use that instantiate the properties which belong to the respective level. Measuring ion conductance at an axon with electrodes, for example, establishes properties on a cytological level; whereas measuring reaction times of participants in a behavioural experiment establishes properties on a psychological level. The way in which different epistemic systems — e.g., a functional magnetic resonance imaging (f) scanner and a blood test — applied on the same entity — a human — establish different properties — a local decrease in de-oxygenated hemoglobin versus, for instance, cortisol levels in the bloodstream — on different levels — the level of brain activation versus the level of endocrine activation — shows that levels and properties are intimately connected. Coming back to our example from the beginning, different properties of the mental state compassion are instantiated in several ways: (a.) through a person as an epistemic system directed towards a myriad of dispositional interactions of her own body, which can then (b.) be picked up as values of standardized questionnaires probing those experiences while, finally, (c.) some properties of the underlying mechanism of compassion are instantiated by means of fMRI (cf. Klimecki et al. 2014). What becomes strikingly apparent in this example is that each of these different ways of property instantiation yields properties we would intuitively base on very different levels. Experience of compassion seems to have a very different quality and complexity to the more abstract numerical values of questionnaires or activation patterns visualized by fMRI. Thus, we can say that the specific way a certain property is instantiated already establishes a corresponding level. This way of defin-

ing levels offers a broader range of application than the levels of mechanisms account, since it is not restricted to properties of mechanisms but rather bears on properties in general. After these considerations, we are now in a position to put down the following as a working definition of levels:

Levels =_{Df} (L) Levels are sets of properties established with respect to their instantiation through the same or a similar kind of epistemic system, which targets the same or a similar dispositional profile as compared to different epistemic systems targeting different dispositional profiles.

Considering the identity criteria of epistemic systems, an epistemic system identical to itself might be involved in the instantiation of two properties, that are on the same level qua being picked up by the said epistemic system—for example, a ruler instantiating the length of 1 meter and 20 centimetres. Two epistemic systems are similar if they pick up exactly the same kind of dispositions and exhibit a similar dimension of output value. For example, two rulers picking up the dispositions of a set of tables to instantiate the height of 1 meter and 59.1 inches (1.5 m) or two humans seeing the color blue. Note that the properties might be different in these cases but they are still on the same level, since they are instantiated by the same or similar epistemic system. A ruler and an infrared detector, for instance, are neither identical nor similar epistemic systems, since they differ in the kind of dispositions they pick up and have different dimensions of output values.

Having provided the two missing definitions for an account of emergence, let us now consider how these can be connected to mechanisms and implemented into a new account of mechanistic emergence. What could properties of mechanisms be? According to the definition of mechanisms given in the introduction, we have to identify entities and activities belonging to the mechanism, as well as starting and termination conditions. These might all be established by property-instantiating epistemic sys-

tems. We decompose, measure, and intervene with the phenomena or their realizing components to establish temporal, functional, or organizational properties. Properties of mechanisms on higher levels are instantiated differently to properties of mechanisms on lower levels. We can use one epistemic system to track particular sequences within a mechanism on a specific level in order to be able to recognize stepwise changes as they unfold in temporal order. For properties on higher levels, however, we have to change the kind of epistemic systems involved. We are dealing with different properties on a different level, and we cannot capture the same causal chain as in the single lower-level mechanism. Instead, we capture a synchronously emergent property on a higher level. A formulation of such an account of mechanistic emergence that incorporates the above definitions of “properties” and “levels” could be constructed as follows:

Mechanistic emergence =_{Df} (ME) Mechanistic emergence is a special form of property instantiation, in which the novelty of the properties is established by a change in epistemic systems involved in their instantiation and through which they span a higher level compared to the level of the components, which realize the higher-level properties by means of their mechanistic organization and process dynamics, thereby changing the overall dispositional profile of the whole while, at the same time, being constrained with respect to their individual dispositions in virtue of the synchronous, non-causal constituency relation.

What is new in this account of emergence is that it acknowledges how even emergent properties ultimately arise out of perfectly explainable mechanistic processes. Unpredictability or unexplainability are no longer the defining characteristics of emergent properties themselves, but only characteristics of the *epistemic context* involved in their instantiation. However, these emergent properties are still *novel* in the sense that they are non-causal, non-aggregative

products of mechanisms that come into existence only on a higher level, established through the kind of property instantiation that none of the components show in isolation.

4 Conclusion

In this commentary my aim has been to point out (i.) that defining levels as crucially dependent on properties has a wider and more flexible range of application than using part-whole relationships as the defining criterion; and to put forward (ii.) that levels of mechanisms and levels of emergence can be reconciled into an account of mechanistic emergence in which the property-dependent definition of levels finds application.

My argument was (a.) that descriptive pluralism, by conceiving of levels essentially as metaphors, cannot yield sufficient conceptual clarification concerning the term “level”—namely, what levels actually are and how they exist and, even undermines the goal of preserving the use of the concept in science. Further, I highlighted (b.) that in “levels of mechanisms”, as presented by Carl Craver, the core criterion for a definition of “level” is a part-whole relationship in conjunction with a constitutive relevance constraint, and that this focuses solely on the vertical dimension existing between levels, and completely omits the more important horizontal dimension of the conditions that must apply for a set of entities to be on the same level. As such, the concept is only very weakly and indirectly characterized, offering little toward its clarification and broader application. Finally, I showed (c.) that ontological novelty is not dependent on epistemic intractability, and that the ontological novelty of properties on higher levels of mechanisms is not a banal fact either in levels of mechanisms or in levels of emergence. What emergence expresses at its core is that new properties are coming into existence and that they are so strikingly novel that they might not be predictable at the moment—or seem to not be so, even in principle. They are novel to such a degree that their instantiation coinstantaneously establishes a wholly new level.

As positive proposals for an alternative view, I defined (d.) “properties” as instantiated by epistemic systems capturing specific dispositional profiles of entities, and (e.) “levels” as sets of such properties instantiated by the same or a similar epistemic system, as compared to those properties instantiated by another epistemic system. Furthermore, I (f.) provided a definition of “mechanistic emergence” implementing the core idea of emergence as aforementioned, together with the proposed definitions of “levels” and “properties”.

Concerning future directions for research, it seems most pressing to further develop the notion of an “epistemic system”. Moreover, the notion of “levels” needs to be further refined with regard to how much epistemic systems can or must differ in order for there to be a new level. Ultimately, of course, it will be intriguing to see whether the developed definitions hold in the light of practical implementations in scientific contexts.

To finish, let us come back to the initial question of whether mind and brain, or more precisely mental processes and neurobiological processes, are on the same level, which we are now in a better position to answer. So far as we acknowledge that mental properties and properties of the brain are different properties, and if we also consider how I defined levels above, we can conclude that mind and brain are in fact on different levels in this sense. But are we thus slipping back into dualism? Absolutely not, since the definition of properties developed above makes clear how it comes about that there can be different properties of one and the same thing: it is dependent on the kind of epistemic system in use to capture specific dispositions or, in short, on the *epistemic context*. Taking up our example of compassion once again, it is now obvious why the phenomenon did not seem to be fully captured only by reference to, for instance, physiological properties. Of course it makes sense to investigate the physiological realization of compassion, to measure autonomic parameters, conduct blood tests, or undertake fMRI scans, but it is equally important to conduct behavioural experiments or even interviews with participants to target the phe-

nomenological experience that encompasses compassion (cf. [Singer & Bolz 2013](#)). This only means that we are doing research on all the different properties of the phenomenon of compassion. We are doing research in different disciplines with different methods, on different levels, and we are capturing different properties of one and the same thing—so let’s work together to incrementally integrate those epistemic contexts and get the complete picture.

Acknowledgements

I wish to express my gratitude to two anonymous reviewers, and in particular the editors, for their extensive comments on an earlier version of this commentary, which contributed immensely to its reaching the present form.

References

- Bird, A. (2007). *Nature's metaphysics: Laws and properties*. Oxford, UK: Clarendon Press.
- Broad, C. D. (1925). *The mind and its place in nature*. London, UK: Routledge & Kegan Paul.
- Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford, UK: Clarendon Press.
- (2015). Levels. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Craver, C. F. & Darden, L. (2013). *In search of mechanisms: Discoveries across the life sciences*. Chicago, IL: University of Chicago Press.
- Ellis, B. (2002). *The philosophy of nature: A guide to the new essentialism*. Chesham, UK: Acumen.
- Hills, D. (2010). Metaphor. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*.
<http://plato.stanford.edu/archives/win2012/entries/metaphor/>.
- Klimecki, O. M., Leiberg, S., Ricard, M. & Singer, T. (2014). Differential pattern of functional brain plasticity after compassion and empathy training. *Social cognitive and affective neuroscience*, 9 (6), 873-879.
[10.1093/scan/nst060](https://doi.org/10.1093/scan/nst060)
- O'Connor, T. & Wong, H. Y. (2009). Emergent properties. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*.
<http://plato.stanford.edu/archives/spr2012/entries/properties-emergent/>.
- Singer, T. & Bolz, M. (Eds.) (2013). *Compassion: Bridging practice and science*. Leipzig, GER: Max Planck Institute for Human Cognitive and Brain Sciences.
- Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind*. Cambridge, MA: Belknap Press of Harvard University Press.

Mechanisms and Emergence

A Reply to Denis C. Martin

Carl F. Craver

I respond to some of the major issues in Martin's commentary, particularly (i) his insistence on a robust notion of being "at" a level, and (ii) his desire for mechanistic emergence to explain the genuine ontological novelty of higher level phenomena.

Keywords

Emergence | Levels | Mechanisms

Author

Carl F. Craver

ccraver@artsci.wustl.edu

Washington University
St. Louis, MO, USA

Commentator

Denis C. Martin

denis.martin@hu-berlin.de

Humboldt Universität zu Berlin
Berlin, Germany

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

The goal of my target essay is to articulate and recommend a view about what one is and is not committed to when one asserts that the domain of neuroscience or some other special science spans multiple levels of organization. How useful one finds my specific recommendation—that they be understood as levels of mechanisms—will depend on the uses for which one deploys the level metaphor. Martin raises a number of questions about my unpacking of the metaphor that help to clarify what is at stake in this discussion.

2 The levels metaphor

First, a word about the idea that the term level is used as a metaphor. According to the Oxford English dictionary, the term "level" derives from terms describing the idea of being parallel to the ground. From there, the term could easily extend to the idea that a building can have multiple levels, or landings, from the ground up. Something like this spatial arrangement of stacked landings seems to be the basis for many applications of the level metaphor (in other words, landings are the "secondary subject" of the metaphor). Usage of the term "level" increased dramatically in the 20th Century, and surely some of this increase is explained by the

fecundity of the metaphor for ordering items in a set from the highest to the lowest (under some understandings of “high” and “low”). These are the specific topics (the primary subjects) to which the secondary subject (landings) is compared. In asserting that the term “level” has this metaphoric aspect, I do not intend to thereby disqualify the metaphor from also expressing a set of true or false commitments about the structure of the world. Indeed, I think that the mechanistic application of the metaphor more or less accurately describes the structure of at least many systems in the domains of the special sciences.

My point in being a pluralist about the levels metaphor is simply to emphasize that the metaphor of a horizontal landing is used to describe a wide variety of altogether distinct primary subjects. As a result, any understanding of how this level metaphor works, what it is committed to, and whether it works for a particular purpose, must begin with a clear understanding of how the metaphor is unpacked in the given application. (One cannot, to pick a perfectly clear example, equate the great chain of being and levels of mechanisms.) A second point was to emphasize that there are many legitimate applications of the metaphor, and that the utility of the metaphor might vary from one application to the next. I am not arguing for a single, correct way of understanding the level metaphor. However, I am arguing that one prominent use of the level metaphor in sciences such as neuroscience can be usefully unpacked as describing mechanistic levels.

3 Being “at” a level

Martin is particularly concerned with a) the idea that there should be a clear sense of what it is to be “at” a level, and b) the idea there is some significant sense in which being at a level represents a genuine ontological novelty.

The first of these concerns arises from the fact that the placement question cannot be (or has not been) given a satisfying answer within the mechanistic application of the metaphor. The mechanistic account of levels focuses, as Martin correctly notes, on what it means to be

at *different* levels of organization within a mechanism. Concerning what it means to be “at a level,” one can say only that two things are “at” the same level just in case they are not at different levels; if neither thing is a component entity/activity in the behavior of the other, then the two things are not at different mechanistic levels and, in this weak sense, they are at the same level. This provides a sort of answer to the placement question, but not one that will satisfy those, such as Martin, who hanker after some additional factor (such as size or similarity) that unites the items at a level and that explains why all such items are at that level. I take the failure to answer the placement question as one of the key revisionary consequences of thinking clearly about levels of mechanisms. It is a crucial guide to understanding what’s misleading about the monolithic conception of levels.

One apparent problem for this consequence (the absence of a satisfying answer to the placement question) was first raised by Lindley Darden (personal communication): X’s -ing might be a component in (and so at a lower level than) S’s ψ -ing, *and* yet both X’s -ing and S’s ψ -ing might be at the same level as (i.e., not at a different level than) some altogether distinct P’s β -ing. There is a failure of transitivity. This, it seems to me, is simply a consequence of the idea that the application of the level metaphor to levels of mechanisms breaks down when one is not talking about relations between parts and wholes. This is unproblematic; it is simply an alternative way of expressing the idea that being “at the same level” is of no additional metaphysical significance within the mechanistic application of the metaphor than simply not being at different levels. This does not prevent one, of course, from using some other ordering criterion for this expressive purpose; one might lump things together on the basis of their sizes or perhaps the instruments used to detect them. But one should be aware that at this point one has left the mechanistic application of the metaphor.

Another consequence of the mechanistic view is that one might equally correctly carve the boundaries of mechanisms in any number of

ways (see e.g., Craver 2004, 2009, 2012). If so, there will not be a uniquely correct answer to the question of how many levels a given mechanism has. Our decisions to privilege some grains in the decomposition of a mechanism as an appropriate place to locate a level depends on our techniques, our theoretical background assumptions, our characterization of the phenomenon, our representational tools, and perhaps certain features of human psychology. As Simon argued, these systems are only *nearly* decomposable; how a system is decomposed depends, for example, on the relative strength of intra- versus extra-system causal interactions that one takes to be appropriate for carving the system at its near-joints. This is another reason I am hesitant to pin too much on the idea of being “at” a level.

4 Ontological novelty and emergence

I am not hostile to the second idea (b) that levels of mechanisms exhibit a kind of ontological novelty, depending on how this novelty is understood and how one thinks that it is achieved. I suggest in the target article that a mechanism as a whole can do things that its parts (taken individually) cannot do. Lawnmowers mow grass; spark plugs do not. But I also claim that a mechanism as a whole cannot do things that its organized and interacting parts cannot do. This is because the behavior of a mechanism as a whole just is (or at least is ontologically intimate with) the organized interaction of its component parts. I don’t know how to make the notion of “ontological intimacy” precise here (though the behavior of a mechanism in context will surely supervene on the organized collection of interacting components in that context). In my hands, “ontological intimacy,” is meant to denote an exhaustive ontological grounding of the behavior of a mechanism as a whole in the organized interactions of its components in a given causal context, however that is properly to be unpacked. Everything has an ontic explanation in terms of the organized activities of the mechanism’s parts.

The term emergence strikes me as suspicious, and I hesitate to use it even in the con-

text of levels of mechanisms, precisely because it suggests a severing of this ontological intimacy, a slide from the banal fact that mechanisms behave as they do because they are organized arrangements of interacting parts to the ontologically esoteric thought that something comes into being with causal powers ontologically inexplicable in terms of the organized interactions of the parts. Such claims about ontological novelty are almost always accompanied by claims that emergent things have a “downward” causal influence. I have labored to drive a wedge between levels of mechanisms and such ideas.

This connects with Martin’s description of Thompson’s view. According to Martin, Thompson requires that the parts in a system must be coupled and dynamically interacting to produce emergent properties and, further, that the properties of the whole should act “downward” on the dynamics of the components (2007). Take these in turn.

The idea of levels of mechanisms, by itself, requires only that the parts be organized and interact with one another; it does not require that the components interact non-linearly. Mechanisms with components that interact non-linearly are a subset of mechanisms more generally. It should not make any difference with respect to the novelty of higher-level properties that the parts interact non-linearly. Such dynamical interactions might make the behavior of the whole harder to predict on the basis of our understanding of the parts; but this is an epistemic, not an ontological, observation irrelevant to the ontological question to which I am responding. To be ontologically novel is not merely to be surprising.

Perhaps systems with dynamically interacting components are harder to idealize into separable interacting components; again, this would appear to be an epistemic issue. Thompson, at least in Martin’s description, is drawing a more fine-grained distinction than I draw. This is perhaps a terminological matter. The pressing question between us is whether non-linearity (or perhaps some other form of complexity) makes any ontological difference to the kind of thing that “emerges” from the “organized interactions” of the parts. I’m inclined

to say no, but a full consideration of the issue would require a more detailed treatment than I can give it here.

Thompson also requires that the higher-level phenomenon should have “determinative” influence on things at lower-levels. “Determinative” surely cannot mean the existence of a universally quantified material conditional with the behavior of the mechanism as a whole as the antecedent and the organized interactions among the components as the consequent. Multiple realization precludes such an analysis. But “determinative” also cannot be understood in the causally productive sense in which something (such as charge or momentum) is passed from whole to part. And for reasons expressed in the target article, there are many widely shared assumptions about the independence of causes and effects that stand in the way of understanding such top-down relations in causal terms. This is why I suggest (as in [Craver & Bechtel 2007](#)) that the language of downward causation is misleading in the context of token mechanisms. We recommend that claims about downward causation are really shorthand ways of expressing an often complex web of constitutive (about relations between things at different levels of mechanisms and levels of realization) and causal claims (about things that are not at different levels from one another) about the system. However this is to be cashed out, the term “determinative” seems misleading.

5 Levels and techniques

Finally, I’ll offer a brief remark concerning Martin’s positive suggestion for reifying levels. Martin’s central idea—that we often pick out things as being on a level because they are detectable with the same or a similar kind of apparatus or with the same or a similar set of procedures—seems to me correct and important. Among the many possible ways of decomposing a spatio-temporal whole into component parts, some of these correspond to items that for one reason or another are readily detectable as such by one or more experimental apparatus. This way of putting things highlights how pragmatic factors, such as available apparatus, determine what we

will take to be an appropriate “landing” in a hierarchy of levels. And it raises useful questions about, for example, when two different apparatus, or two different tasks, or two different procedures in fact target the same items, or items at the same level. On a fine-grained characterization of our experimental instruments, no two experiments are the same. On a coarse grain, even superficially quite distinct experiments can be targeting the same phenomenon (consider, e.g., implicit bias or spatial memory). As noted above, it seems to me that many other epistemic, theoretical, and psychological factors enter into these decisions as well.

This good point, however, is obscured by an ontology in which properties apparently come into existence during acts of detection. Martin’s view is that objects have disposition profiles, and these profiles are turned into properties when they are measured. It is unclear to me, however, why one would not want to say that properties are there to be detected all along, or perhaps that properties just are the dispositional profiles of things. Martin doesn’t do much to motivate this experimental idealism about properties, but the thought seems hard to motivate, at least for macroscopic phenomena. It is an apparent consequence of Martin’s account that levels don’t exist until they are detected. And it is an apparent consequence of his view that new levels come into existence when we develop new instruments to detect them. And, to reiterate the thought in the last paragraph, we will have to wrestle with the question of when two techniques detect the same thing or different things. Martin’s positive proposal makes this question much more pressing, given that, for Martin’s view, the structure of the world—specifically, the distribution of properties in space and time—apparently hangs in the balance. But if we abandon the idea that the placement question must have a uniquely correct answer, these questions are less pressing for thinking about ontology; nonetheless, questions about how we coordinate different experimental tasks, protocols, and procedures are at the heart of the epistemological challenge faced by any experimentalist (see [Sullivan 2009, 2010](#)).

References

- Craver, C. F. (2004). Dissociable realization and kind splitting. *Philosophy of Science*, 71 (5), 960-971.
- (2009). Mechanisms and natural kinds. *Philosophical Psychology*, 22 (5), 575-594. [10.1080/09515080903238930](https://doi.org/10.1080/09515080903238930)
- (2012). Functions and mechanisms: A perspectivalist account. In P. Huneman (Ed.) *Functions* (pp. 133-158). Berlin, GER: Springer.
- Craver, C. F. & Bechtel, W. M. (2007). Top-down causation without top-down causes. *Biology and Philosophy*, 22 (4), 547-563. [10.1007/s10539-006-9028-8](https://doi.org/10.1007/s10539-006-9028-8)
- Sullivan, J. (2009). The multiplicity of experimental protocols: A challenge to reductionist and non-reductionist models of the unity of neuroscience. *Synthese*, 167 (3), 511-539. [10.1007/s11229-008-9389-4](https://doi.org/10.1007/s11229-008-9389-4)
- (2010). Reconsidering spatial memory and the Morris water maze. *Synthese*, 177 (2), 261-283. [10.1007/s11229-010-9849-5](https://doi.org/10.1007/s11229-010-9849-5)
- Thompson, E. (2007). *Mind and life: Biology, phenomenology, and the sciences of the mind*. Cambridge, MA: Belknap Press of Harvard University Press.

Mental States as Emergent Properties

From Walking to Consciousness

Holk Cruse & Malte Schilling

In this article we propose a bottom-up approach to higher-level mental states, such as emotions, attention, intention, volition, or consciousness. The idea behind this bottom-up approach is that higher-level properties may arise as emergent properties, i.e., occur without requiring explicit implementation of the phenomenon under examination. Using a neural architecture that shows the abilities of autonomous agents, we want to come up with quantitative hypotheses concerning cognitive mechanisms, i.e., to come up with testable predictions concerning the underlying structure and functioning of an autonomous system that can be tested in a robot-control system.

We do not want to build an artificial system that is, for example, conscious in the first place. On the contrary, we want to construct a system able to control behavior. Only then will this system be used as a tool to test to what extent descriptions of mental phenomena used in psychology or philosophy of mind may be applied to such an artificial system. Originally these phenomena are necessarily defined using verbal formulations that allow for interpreting them differently. A functional definition, in contrast, does not suffer from being ambiguous, because it can be expressed explicitly using mathematical formulations that can be tested, for example, in a quantitative simulation. It is important to note that we are not concerned with the “hard” problem of consciousness, i.e., the subjective aspect of mental phenomena. This approach is possible because, adopting a monist view, we assume that we can circumvent the “hard” problem without losing information concerning the possible function of these phenomena. In other words, we assume that phenomenality is an inherent property of both access consciousness and metacognition (or reflexive consciousness). Following these arguments, we claim that our network does not only show emergent properties on the reactive level; it also shows that mental states, such as emotions, attention, intention, volition, or consciousness can be observed, too. Concerning consciousness, we argue that properties assumed to partially constitute access consciousness are present in our network, including the property of global availability, which means that elements of the procedural memory can be addressed even if they do not belong to the current context. Further expansions are discussed that may allow for the recognition of properties attributed to metacognition or reflexive consciousness.

Keywords

Access consciousness | Consciousness | Emergent properties | Internal body model | Minimal cognitive system | Motor control | ReaCog | Recurrent neural networks | Reflexive consciousness | Robotic architecture

1 Introduction

In this article we propose a bottom-up approach to higher-level mental states, such as, for example, consciousness. In contrast to most related approaches, we do not take consciousness as our point of departure, but rather aim,

firstly, to construct a system that has basic properties of a reactive system. In a second step, this system will be expanded and will gain cognitive properties in the sense of being able to plan ahead. Only after this work is finished, we

Authors

[Holk Cruse](#)

holk.cruse@uni-bielefeld.de
Universität Bielefeld
Bielefeld, Germany

[Malte Schilling](#)

malte.schilling@uni-bielefeld.de
Universität Bielefeld
Bielefeld, Germany

Commentator

[Aaron Gutknecht](#)

aaron-gutknecht@gmx.de
Goethe-Universität
Frankfurt a. M., Germany

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu
Monash University
Melbourne, Australia

ask to what extent this system is equipped with higher-level properties as for example emotions or consciousness. While other approaches would require an exact definition of, for example, consciousness, in our case we do not have to start from a clear-cut definition and try to fit it into a model. We follow this alternative route because there are no generally accepted definitions concerning these higher-level phenomena. In this way we hope to identify the essential elements required to instantiate, for example, consciousness.

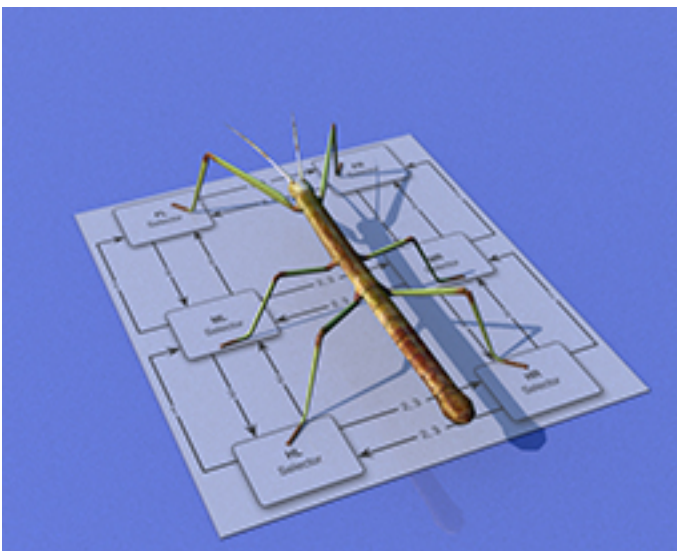


Figure 1: Arrangement of the leg-controllers (boxes: FL front left, ML middle left, HL hind left, FR front right, MR middle right, HL hind right) of the hexapod walker. The arrows show coordinating influences (1–4) that act between neighbouring leg-controllers.

The idea behind this approach is that higher-level properties may arise as emergent properties, i.e., may occur without requiring explicit implementation of the phenomenon under examination but instead arise from the cooperation of lower-level elements. Some authors distinguish between “strong” emergence and “weak” emergence (e.g., Laughlin & Pines 2000). Strong emergence means that there is principally no way to explain the emergent property by known properties of the elements of the system and their coupling. Here we are dealing with weak emergence. In this case, a property recognized when looking at the whole system can at first glance not be traced back

(or perhaps only partially) to known properties of the elements and their couplings. Often, auxiliary assumptions are made to explain this property as a global property, i.e., as a property ascribed to the system as a whole. A more detailed inspection may, however, show that such auxiliary assumptions are not required. Instead, the emergent property follows from the properties of the elements and the specific ways in which they causally interact. This insight allows for an understanding of an emergent property in the sense that this property can be predicted, although we may not understand why it arises, and that one is able to construct a new system showing this property.

Following this approach, one crucial decision to be made at the beginning concerns the granularity of the lower-level elements. In our approach, we start from a behavioral perspective and focus on the nervous system as central to the control of action. Therefore, we use neuronal units as the basic elements for our modeling and for the analysis. Specifically, we use artificial neural network units with analogue activation values and dynamic (low-pass filter) properties¹. That is, our neural elements are qualitatively comparable with non-spiking neurons. Although there are arguments that consciousness, in order to arise, might require synchronously oscillating spikes (Singer & Gray 1995), we claim that the level applied here is general enough to allow for an understanding of such mental processes. As a side effect, this level of abstraction covers different evolutionary groups, such as those represented by insects and mammals, for example. Though much of our discussion, below, focuses on the example of insects, we do not want to argue that insects have all the higher-level properties addressed later in this article, but only that they share the same fundamental functions used in motor control and have, on that level, a comparable structure.

Using these simple neural elements, we start by implementing very basic faculties that include the ability to move one’s own (non-

¹ A low-pass filter is qualitatively characterized by an increase of output activation that, when excited by a constant stimulus, asymptotically approaches a given output value.

trivial²⁾ body, and allow for orientation and navigation in an only partially known environment. To this end we use a body with six, insect-like legs. This means that we deal with at least eighteen active degrees of freedom (DoF) and not two—as is the case for many robots that are restricted to moving around on a two-dimensional plane. This means that the controller has to deal with a large number of redundant DoFs. To control the behavior of the robot we use a reactive and embodied neuronal controller, as it is available from earlier work on insect behavior (Schilling et al. 2013a). Later, a minor expansion of the network will allow for cognitive faculties.

on the walking behavior of stick insects (Dürr et al. 2004; Bläsing 2006; Schilling et al. 2013b). As will be explained in section 2, Walknet was set up as a system for controlling the walking behavior of a six-legged system in an unpredictable environment, e.g., on cluttered terrain or climbing over large gaps—which, when performed in a realistic, natural environment is a non-trivial task. Already on this level we can observe emergent properties. The number of legs on the ground differs depending on the velocity of the walker (for slower walking more legs are on the ground). As a consequence the phase relations between different legs differ depending on the velocity of the walker. Importantly, the resulting stepping patterns (“gaits”) are not explicitly encoded in the control network, but are a result of the interaction of the control network with the environment as mediated through the body (1st order embodiment Metzinger 2014). In a further step, the reactive

Cruse, H. & Schilling, M. (2015). Mental States as Emergent Properties - From Walking to Consciousness. In T. Metzinger & J. M. Windt (Eds). *Open MIND*: 9(C). Frankfurt am Main: MIND Group. doi: [10.15502/9783958570436](https://doi.org/10.15502/9783958570436)

controller is expanded to be able to deal with navigation tasks. This additional network, called “Navinet”, is able to simulate a number of experimental results observed in desert ants and honeybees, such as the capability of finding food sources using path integration and orientation with respect to visual landmarks.

Both networks are characterized by their decentralized nature. These networks consist of procedural, (reactive) elements, namely small neural networks that in general connect sensory input with motor output, thereby forming the procedural memory. Inspired by (Maes 1991), these procedural elements are coupled via a “motivation unit network”, a recurrent neural network (RNN) that forms the backbone of the complete system. This type of architecture has been termed MUBCA (for Motivation Unit Based Columnar Architecture (MUBCA), Schilling et al. 2013b). The motivation unit network allows for selection of different behaviors by adopting different attractor states, where each attractor represents a group of motivation units being activated, which in turn control the procedural elements. As the different groups do in part overlap, albeit in different ways, the network allows for the representation of a heterarchical structure (e.g., see left upper part of figure 2).

As a next “evolutionary” step, this reactive network will be expanded to be able to embrace cognitive properties (sects. 3 and 6). The notion of cognition is often used in a broad and sometimes unspecific way. In the following we will rely on the definition given by McFarland & Bösner (1993) who assume that *a cognitive system is characterized by the capability of planning ahead*. We prefer this clear-cut definition of cognition compared to many others found in the literature, as the latter are generally quite weak (in extreme cases cognitive properties are even attributed to bacteria, which, in our view, would make the term cognition meaningless). While such a specific definition might seem too narrow, in our understanding it captures the essence of cognition. Focusing on planning ahead being realized by mental simulation (Hesslow 2002) allows extending this notion of cognition to easily include other high-level phenomena,

while still relying on the same internal mechanism. Therefore, in this article, apart from section 10.3 (Metacognition) we will use the term cognition in the strict sense as proposed by McFarland & Bösner (1993).

Being able to plan ahead implies the capability of being able to internally simulate behavior, which basically means to be able to simulate movements of one’s own body within a given environment. This faculty requires, as a first step, the availability of a flexible, “manipulable” internal body-model. Planning ahead is interesting in a situation where the actually carried out reactive behavior cannot reach the currently pending goal. Therefore, a further expansion is required that allows for the invention of new behaviors. Together with the faculty of planning ahead, the system can then test newly-invented behaviors by applying internal simulation (“internal trial-and-error”) in order to find a solution for novel problems for which no solution is currently known to the system.³

This system, called “reaCog”, represents a basic version of a cognitive system in the strict sense intended by McFarland & Bösner (1993). As such, cognitive expansion does not function by itself, but only, like a parasite, on top of the reactive structures—a view that has been supported for a long time (Norman & Shallice 1986). The cognitive system depends on its reactive basis (therefore it is called reaCog). Therefore, the evolution of cognitive abilities crucially requires a correspondingly rich (procedural) memory.

In order to increase the richness of the memory of the complete system, in section 5 we introduce perceptual memory and complete the system by implementing “Word-nets”, a specific form of procedural and perceptual memory. In this way, the whole system is equipped with aspects of semantic memory, and can be claimed to represent a minimal cognitive system. We do not deal with learning but only discuss the properties of the finished network. The learning

³ Note that the term simulation is used here in two different ways. “Internal simulation” enables the agent to simulate behaviors internally, i.e. without actually performing them in reality. Simulation of an animal addresses the construction of an artificial agent. The agent may take the form of a software simulation or a hardware simulation (i.e., a physical robot).

of some aspects has, however, been treated earlier (Hoinville et al. 2012; Cruse & Schilling 2010a).

After having introduced reaCog in sections 2–6, we will, in sections 7–11, discuss how more abstract functions, such as those described in, e.g., psychology, can be based on such a simply-structured network.

A fundamental problem when aiming for an understanding of phenomena like emotions or consciousness concerns the phenomenal aspect. The phenomenal aspect, often characterized as the hard problem (Chalmers 1997), refers to the strange, unexplainable phenomenon that physical systems, in our case represented by specific dynamics of neuronal structures, can be accompanied by subjective experience. Basic examples are experiencing pain, a color, or the internal state of an emotion (e.g., joy, fear). In section 7 we discuss this aspect in some detail and postulate that phenomenality is an emergent property. As mentioned, we are not aiming to solve the “hard” problem (Chalmers 1997), but we argue that it is sufficient to concentrate on the functional aspect.

In particular, we focus on the phenomena of emotions and consciousness. According to a number of authors (e.g., Valdez & Mehrabian 1994), these are assumed to be an inherent property for some cognitive systems. Therefore, although we do not want to state that emotions (section 8), attention, volition, intention (section 9), and consciousness (section 10) should necessarily be attributed to our system in any sense, we want to discuss to what extent properties characterized by different levels of description can be observed in our model.

Considering emotions, these are defined on different levels in the literature, so that there is no clear, generally accepted distinction between concepts like emotions, moods, motivations, drives, etc., which appear to form a continuum of overlapping, not clearly separable concepts (Pérez et al. 2012). Focusing on selected examples, in section 8 we will show how these phenomena may be attributed to our system, for example by referring to basic emotions as proposed by Ekman (1999).

Concerning consciousness, as discussed by Cleeremans (2005), this phenomenon should be approached by differentiating different aspects and treating those aspects separately. To this end, following Block (1995, 2001), Cleeremans (2005), introduces a distinction between access consciousness, metacognition, and phenomenal consciousness. In sections 10.1 (access consciousness) and 10.3 (metacognition) we will focus on whether and how the presented model can be related to the different aspects that are described by Cleeremans (2005), such as access consciousness and metacognition. From our point of view the simple control structure presented does fulfill some aspects of both access consciousness and metacognition. We shall finish with discussion and conclusion in sects. 11, 12.⁴

2 Walknet

ReaCog is an expansion of a control system that has been realized as a neural network. The underlying system has been termed Walknet. Walknet is biologically inspired and is supposed to describe the results of many behavioral studies on the walking behavior of stick insects (Dürr et al. 2004; Schilling et al. 2013b). We will briefly sketch the properties of the network as far as is required for understanding the basic abilities considered here.

Overall, the controller has to deal with the difficult task of coordinating multiple degrees of freedom; in the case of the hexapod walker the body consists of twenty-two DoF. There are three DoF for each of the six legs and an additional four DoF are present in between the body segments. The system is redundant, as only six DoFs are needed to define a position and orientation in three-dimensional space. The controller therefore has to deal with sixteen extra DoFs. The architecture of the Walknet controller is decentral. Each leg has an individual and more or less independent controller that decides which action to choose (two such leg-controllers are shown in figure 2, the black boxes in the lower part). A single leg

⁴ This article comprises an essential extension of an earlier paper (Cruse & Schilling 2013).

controller consists of several procedures. In the figure, each procedure is represented as a single black box. In the basic system, the two important behaviors a leg can perform are the swing and stance movement. The procedures themselves are realized as artificial RNN. Examples are the two basic procedures: the “Swing-net”, which controls the swing movement, and the “Stance-net”, which controls the stance movement of the leg. Only two of the six leg-controllers are shown. These networks constitute the procedural memory of the system. The procedural modules receive direct sensory input and provide motor control commands as an output. But there are also modules that provide input to another module. The controller on the leg level determines which procedure should be activated at any given time, depending on the current state of the leg (swing or stance), as well as on sensory inputs (ground contact, position). In addition, controllers of neighboring legs can influence each other through a small number of connections between those controllers. These influences are explicitly derived from experiments on the coordination of legs in walking experiments on the stick insect.

As was found in the insects, during the swing movement (protraction) the legs aim towards a position at the front, close to the position of the anterior leg. Therefore, each leg possesses a so-called “target net” in order to produce these targeted movements. During forward walking the so-called “Target_fw-net” is responsible for this targeting. During backward walking “Target_bw-net” is used. Both directly influence the Swing-net. Procedures marked as blue boxes (“body model”, “leg model”) will be explained below (section 3.1).

ReaCog is expanded by an RNN, which consists of motivation units (figure 2, marked in red). This network allows the system to autonomously select one of the different possible behaviors. For example, the system may choose between forward or backward walking, or standing. A motivation unit is an artificial neuron with linear summation input and piecewise linear activation function, showing output values from zero to one. Applied to a

procedure, for example Swing-net, a motivation unit determines the strength of the output of the corresponding procedural network (in a multiplicative way). As mentioned above, motivation units form a recurrent neural network and can influence each other through excitatory or inhibitory connections (as shown in figure 2).

In addition, there are sensory units that are part of this RNN and that can directly influence the motivation units’ activation, e.g., as shown in figure 2 for the “lower-level” units for Swing and Stance. There, an active ground-contact sensor of a leg reinforces the stance motivation unit for this leg. As the motivation unit network can be arbitrarily expanded, it allows to control of complex behaviors. To illustrate a small group of behaviors only, units as “walk”, “fw” (forward), “bw” (backward), “leg1” are depicted (for more examples see Schilling et al. 2013b; Cruse & Wehner 2011).

The network of motivation and sensory units does not have to form a simple, tree-like structure (see figure 2). It can constitute a heterarchy. Motivation units can be bi-directionally connected through positive (arrow-heads) and negative (T-shaped heads) connections. As shown in the figure, this can lead to cycles. There are also different overlapping subnetworks, e.g., the “leg” units as well as the motivation unit for “walk” are active during backward and forward walking. But only one unit indicating the direction of walking can be active at any given time, i.e. either the unit “fw” or “bw” can be active. As a consequence, there are multiple stable attractor states formed through the combinations of excitatory and inhibitory connections. The stable “internal states” stabilize the behavior of the overall control system, as the system cannot be easily disturbed solely through inappropriate sensory inputs. For example, sensory inputs are treated differently depending on the current state (swing or stance) of the control system, and these internal states can be differentiated on a higher-level, e.g., into walking, standing, or feeding (for details see Schilling et al. 2013a; Schilling et al. 2013b).

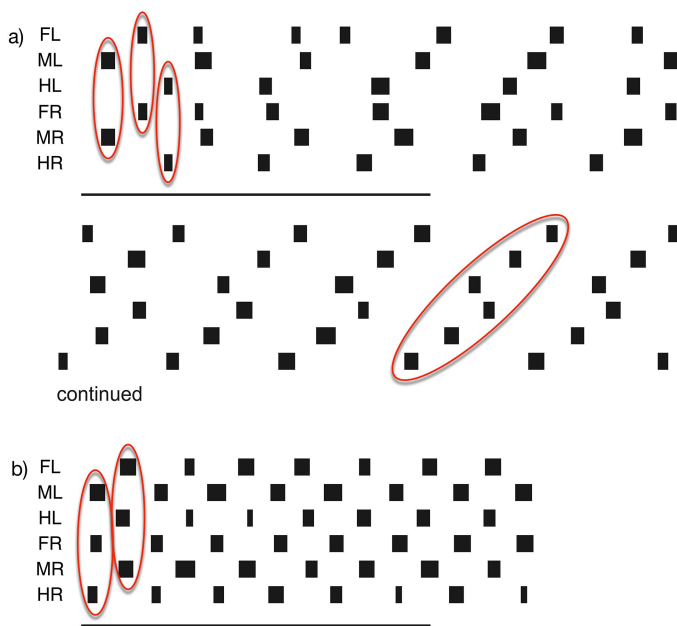


Figure 3: Step pattern arising from the decentralized leg-controllers connected by local rules and the environment. Abscissa is time; black bars indicate swing movement; the gaps represent stance movement of this leg (from top to bottom: front left leg (FL), middle left leg (ML), hind left leg (HL), correspondingly front right leg (FR), middle right leg (MR) and hind right leg (HR) for the right side). The lower bars indicate 500 iterations corresponding to 5s real time. These “foot-fall patterns” show various locally or globally stable patterns depending on walking velocity (a: slow, b: fast) and of starting position. In (a) the legs start with an “uncomfortable” leg configuration leading to a gallop-like pattern (indicated by the vertical ellipses) that after about six steps changes to the globally stable pattern, typical for slow insect walking (see inclined ellipses, step # 8). (b) shows fast walking leading to a tripod gait characterized by synchronous swing movements of ML, FR, HR and FL, HL, MR (see vertical ellipses).

For an RNN, maintaining a stable state is a non-trivial problem, in particular, when there are various disturbances. To illustrate the adaptability and at the same time the stability of the behavior controlled by such a motivation unit network, in figure 3 we show two cases of hexapod walking. Figure 3a shows an example of a slow walking speed where the legs begin from a difficult starting configuration (both front legs, both middle legs and both hind legs start from the same position, which is opposite to the coordination found in normal walking,

where opposite legs alternate). Nonetheless, the agent is able to walk. After some steps, the agent reaches a temporally stable pattern corresponding to normal walking. Figure 3b shows a step pattern corresponding to high-speed walking, often termed “tripod gait”. Although usually considered to be a regular pattern, detailed inspection shows that there are local temporal variations, but the overall pattern remains stable (for videos of further walking examples see Schilling et al. 2013b). It is important to note that none of these step-patterns are explicitly implemented, but arise as emergent properties (for details see Schilling et al. 2013a). As another impressive emergent property, Bläsing (2006) showed that, with some minor extensions, this walker is able to climb over large obstacles (which can be more than twice the normal step-width).

3 Internal representation

In addition to using the loop through the environment itself, some form of internalization is a prerequisite for any kind of planning. Therefore, specific internal representations⁵ are necessary for a cognitive system. This is well in line with the embodied perspective, because from an evolutionary point of view internal models are not at first disconnectable from a very specific function, and they work in service of a specific behavior (Glenberg 1997). Internal models have, in this sense, co-evolved with behavior (Steels 2003). An early representation is the representation of one’s own body, and such a representation becomes meaningful early on, in simple control tasks like targeted movements or sensor fusion.

3.1 Body model

In reaCog we introduced an internal model of the body. This model is realized as an RNN (Schilling 2011) and has a modular structure (Schilling & Cruse 2007; Schilling et al. 2012). The overall model consists of two different

⁵ The term representation is used here in the broad sense of Steels (1995) “physical structures (for example electro-chemical states) which have correlations with aspects of the environment”.

levels. On the top level the whole body and the structure of the insect are represented in an abstract way. Only on the lower level are the details filled in. The lower level consists of six leg networks. Here, for each leg the functional structure of the joints and the limb is captured. In this way this level of representation can be used for motor control and provides detailed information about joint movements. On the higher level, the structure of the body and the legs is represented in an abstract form, i.e., only the footholds of the legs appear on this level. Figure 2 shows the different parts of the body model (drawn in blue). The body model is modular. It comprises a holistic system that is realized as an RNN (figure 5, see Schilling 2011; Schilling et al. 2012 for details).

The body model is used during normal walking, meaning that the system is still in the reactive mode, in forward as well as backward walking or when negotiating curves. It coordinates the movement of the joints and delivers the appropriate control signals for the Stance-networks. As explained above, overall the system is redundant, with twenty-two DoFs in the whole body structure, and this makes deriving consistent control signals for all the joints a difficult problem that can't be computed directly, but rather requires application of additional criteria (e.g., for optimizing energy consumption). In our approach, which uses the internal body model, we employ the passive motion paradigm (von Kleist 1810; Mussa-Ivaldi et al. 1988; Loeb 2001). Consider the body model as a simulated puppet of the body (figure 5) that is pulled by its head in the direction of the goal (figure 5b, pull_fw). This information on the target direction could be provided by sensory input, e.g., from the antennae or vision, in the form of a target vector (figure 2, sensory input). When pulled in this direction, the whole model should take up this movement and therefore the individual legs currently in stance should follow the movement in an appropriate way. The induced changes in the joints can be read out and applied as motor commands in order to control the real joints. In backward or curved walking, the body model has only to be pulled into a corresponding direction (in backward walking

using the vector attached to the back of the body model, pull_bw (figure 5b). In this way we obtain an easy solution to the inverse kinematic problem as the body-model represents the kinematical constraints of the body of the walker. It restrains the possible movements of the individual joints through these constraints, and only allows possible solutions for the legs standing on the ground, thereby providing coordinated movements in all the involved joints.

The body-model is also connected to the sensors of the walking system and integrates the incoming sensory information into the currently-assumed state of the body as represented in the body-model. In this way the body-model is able to correct noisy or incorrect sensory data (Schilling & Cruse 2012). Overall, the main task of the body model is pattern completion. It uses the current state and incoming sensory data to come up with the most likely state of the body that fulfils the encoded kinematic constraints. In this way, the model can also be used as a forward-model, meaning that, given specific joint configuration, the model can predict the three-dimensional arrangement of the body, for example the position of the leg tips. The predictive nature of the model is crucial as it allows exploiting the model for planning ahead (see below). It is important to note that while we do not want to claim the existence of such a model in insects, the functions of internal models are prediction, inverse function, and sensor fusion, and these can all already be found in insects.

3.2 Representation of the environment

Of course, internal representation should also contain information on the surroundings. We started with a focus on the body and want to extend this network in a way that reflects how the environment affords (Gibson 1979) itself to the body, i.e., a focus on interaction with the environment.

As an example of how the reaCog architecture could be extended to include representation of meaningful parts of the environment, we want to briefly sketch an expansion of Walknet that would allow for insect-like navigation ("Navinet" Cruse & Wehner 2011; Hoinville et

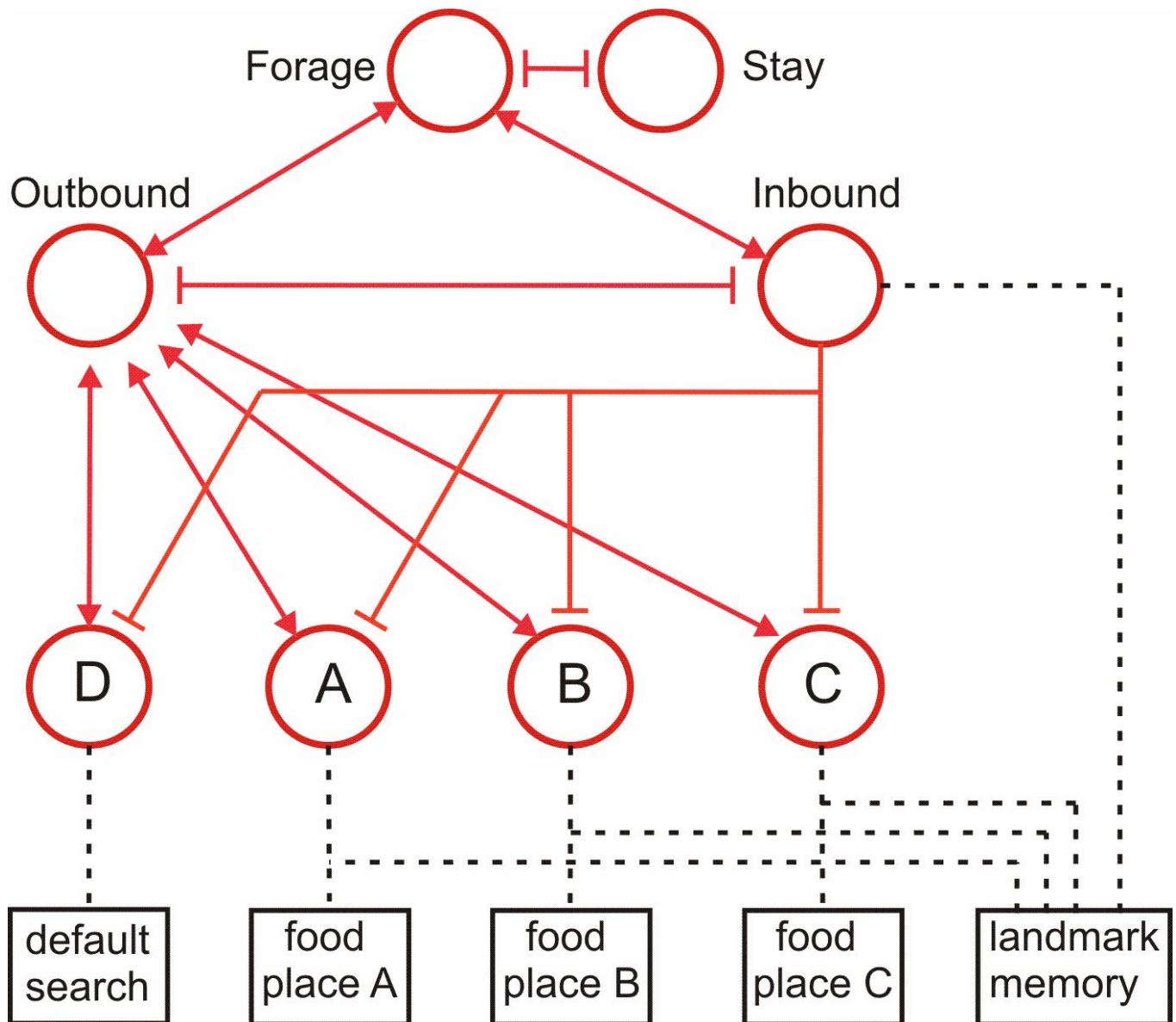


Figure 4: Motivation unit network of Navinet for the control of ant-like navigation. Unit Outbound controls travel from the home to a food source (A, B, C) or a default search for a new source (D). Unit Inbound controls travel back to the home. Memory elements (black boxes) contain position and quality of the food source (A, B, C) or information on visual landmarks (landmark memory).

al. 2012). Navinet provides an output that will be used by the body-model explained above to guide walking direction. Due to the network, the agent can make an informed decision about which learned food source she will visit (e.g., sources A, B or C), or if she is travelling back home or not (Outbound, Inbound, respectively). The output of Navinet is, in this way, on the one hand tightly coupled to the control of walking and the representation of the body. On the other hand, Navinet is constructed using motivation units in the same way as the walking con-

troller, and those motivation units take part in the action-selection process. Importantly, Navinet (like desert ants) shows the capability of selective attention, since it is context dependent and only responds to learned visual landmarks in the appropriate context, i.e., when related to the current active target food source. The structure of the motivation-unit network is sketched in figure 4. Examples of possible stable internal states are (Forage – Outbound – source A – landmarks associated with source A) or (Inbound – landmarks associated with Inbound),

for instance. As an interesting emergent property, Navinet does not presuppose an explicit “cognitive map”. Such a map-like representation has been assumed necessary by several other authors (Cruse & Wehner 2011). How learning of food source positions and food quality is possible has been shown by Hoinville et al. (2012).

4 Planning ahead, cognition

Even though Walknet is set up as a fixed structure consisting of hard-wired connections of the RNN, it can flexibly adapt to disturbances in the environment as needed during, for instance, crossing large gaps (Bläsing 2006). Nonetheless, the system might of course run into novel situations that require an even higher degree of adaption, and as such will require novel behaviors. As an example, think of a situation in which all the legs except the right hind leg are in the anterior part of the working range. When the right hind leg is forced to lift from the ground as it approaches a position very far to the rear, the whole system will become unstable, as the center of gravity is positioned very far towards the rear of the animal. In this case, the center of gravity would not be supported by the other legs, nor by the right hind leg that tries to start a swing movement. As a consequence, the agent would fall over, backwards. This problem could be detected by “problem detectors”, e.g., specific sensory input that reacts to the specific load distribution (a different solution is explained in section 8). In order to overcome this problem, the system would have to break out of its usual pattern of behavioral selection and try to select a different behavioral module that is usually not applicable in the given context. For instance, making a step backward with the right middle leg would be a possible solution, as this would provide support for the body and would afterwards allow going back to the normal walking behavior and the subsequent swing movement of the right hind leg. Usually, backward steps can only be selected in the context of backward walking.

Figure 6 shows an expansion that allows the system to search for solutions that are not connected to the current context. This expansion

is termed the “attention controller”. We introduce a third layer of units (figure 6, in green), that is essentially a recurrent winner-take-all network (WTA-net). For each motivation unit there is a corresponding partner unit in this WTA-network. Currently-active motivation units suppress their winner-take-all (WTA) partner units (T-shaped connections in figure 6). Therefore, a random activation of this WTA-net will lead to the activation of one single unit not belonging to the currently-activated context. The random activation will be induced by another parallel layer, the “Spreading Activation Layer” (not depicted in figure 6, further details are described in (Schilling & Cruse submitted)). The winning unit of the WTA layer then activates its corresponding motivation unit. This triggers the connected behavior that can be tested as a solution to the problem at hand. The network follows a trial-and-error strategy as observed in, e.g., insects.

As has been proposed (Schilling & Cruse 2008), a further expansion of the system that is, most probably, not given in insects is not the testing of a behavior in reality, but instead the application of a newly-selected behavior on the body-model and the use of the model instead of the real body. The motor output is routed to the body-model instead of to the real body, and the real body is decoupled from the control system while testing new behaviors. Due to the predictive nature of the body-model, it can be used to predict possible consequences and to afterwards decide if a behavior solves the current problem and should be tried out on the real body. This procedure is called internal simulation and requires the introduction of switches that reroute motor output signals from the real body to the body model (figure 6, switch SW). Only after a successful internal simulation will the behavior be applied to the real body. McFarland & Bösner (1993) defined a cognitive system as a system that has the ability of planning ahead, i.e., that is able to perform internal simulations in order to predict possible outcomes of behaviors. Therefore, this latter expansion would make the control system cognitive (for details see Cruse & Schilling 2010b).

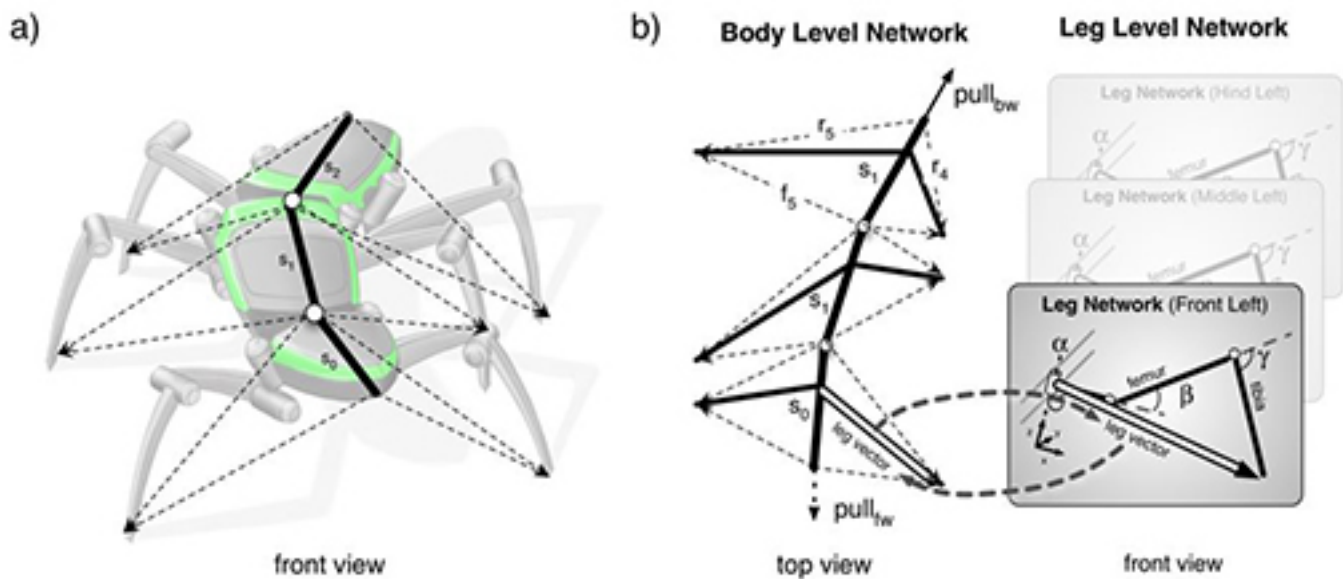


Figure 5: The body-model and its relation to the body of robot Hector (a). (b) shows the vectors forming the central body (left) and the vectors forming one leg model (right). The central model and the leg-models are connected via the shared “leg vector” (white arrows) that point from the hip to the tip of the leg (shown here for the left front leg only). Walking direction and velocity are controlled by the input vectors pull_fw (forward) or pull_bw (backward) provided by sensory input.

5 Word-net and perceptual memory

In our network, we have up to this point only dealt with procedural memories, i.e., memories representing the connections between specific sensorimotor elements that are able to control specific behaviors (e.g., Swingnet, landmark). As a final extension, we will now show how the network might also be equipped with some aspect of semantic memory, such that meaning can be attributed to verbal expressions. To this end, the network can be expanded through the introduction of another layer (not shown in figure 6). In this fourth layer, verbal expressions are stored as procedures or “Word-nets”. These procedures can either be used to pronounce a stored word or to comprehend it, i.e., they can be used for motor control and for auditory perception. As is the case for other procedures, each Word-net is equipped with a motivation unit. As the motivation units of Word-nets have a specific function, for an easier distinction we will call them word units (WU). Following Steels (2007; Steels & Belpaeme 2005) each Word-net is related to a corresponding unit of the motivation network that carries meaning

(e.g., the motivation unit for walking is connected to a Word-net “walk”). The meaning of the Word-nets is in this way grounded in the behaviors of the corresponding motivation units. As an example, figure 7 shows a possible detail of such a network, including some elements of Walknet and Navinet. The motivation units of a procedure (e.g., Swing net) and its corresponding Word-net (e.g., “Swing”) are coupled via bi-directional connections (dashed double-headed arrows). The connections cannot be active at the same time, but depend on an overall state of the network, termed “Report” and “Perceive”. In the Perceive state, only connections from the word unit to the motivation unit of its non-word procedure can be activated (from top to bottom in figure 7), whereas in the Report state only the opposite connections can be activated. As can be seen in figure 7, Word-nets can not only be connected with motivation units of the sensorimotor nets, but also with motivation units that do not directly control a sensorimotor element (e.g., Walk, Outbound).

What might be the function of this extension by Word-nets? In the Perceive state (or react state), a perceived word, uttered by another

agent, will activate, via its word unit, its partner's motivation unit, and thereby possibly influence behavior (depending on the actual internal state of the system and on the strength of the word input). When in the Report state, the actually active motivation units will in turn activate their corresponding word units, which may lead to an uttering of a word. As, of course, only one word can be activated at a given time, some kind of decision network (e.g., a WTA net) is required, though, for reasons of simplicity, not shown in figure 7. In any case, introduction of Word-nets allows for a very basic form of communication between the agent and any other partner, communication being limited to "one-word sentences".

As indicated on the left side of figure 7 (units "front", "left"), further motivation units might be introduced into the network that do not have a direct function within, in this case, the Walknet controller. Of course, these units may be connected to word units. (Note that we do not deal with the question how these units may be connected within the network through training).

This architecture combines sensorimotor procedures with Word-nets (which by themselves represent specific sensorimotor procedures). Together, they form a simple case of semantic memory, because procedural memory representing an action (e.g., Swing-net) is connected with a memory element representing verbal symbols.

To illustrate the versatility of this architecture, we will briefly address how it can also be applied in order to embrace perceptual memory. Following ideas of O'Connor et al. (2009), Cruse & Schilling (2010a) have shown how an RNN, using the same elements as applied here for the motivation unit network, could be used to construct a perceptual memory. This network does not only allow the representation of directly perceived perceptual elements (e.g., the colour or shape of an object), but also of superordinate concepts (e.g., Cow, Animal, four-legged). Note that "four-legged" might also be a feature of non-animals, e.g., a table. Therefore, the ability of our network to deal with heterarchical structures is advantage-

ous for perceptual memory, too. Elements of such a distributed memory can also be connected to specific Word-nets (e.g., "red", "Cow", "animal"), as has been explained above for the sensorimotor motivation units. Correspondingly, activation of one memory element of this perceptual memory may elicit the uttering of the corresponding word, and, in turn, when in Perceive mode, the hearing of a word may activate various elements of the procedural memory that are associated with this word.

6 ReaCog: Emergent properties characterized by applying other levels of description

To summarise, the neural controller Walknet, (for details see Dürr et al. 2004; Schilling et al. 2013a) is an embodied control system (first-order embodiment, cf. Metzinger (2006, 2014)). The reactive system can deal with varying unpredictable environments. It relies only on information that is available to the given mechanosensors, which is possible because both body and environment are integral to the overall computational system. In this way, the system is embodied. Of course, the system has a physical body, but even more, being embodied means that properties of the body (like its geometry) are exploited in computations of the controller. Using its own body as part of a loop through the world allows for dramatically simplifying computations (Schmitz et al. 2008). These properties are of course also present in the expanded version, reaCog. Even though in reaCog an internal body-model is introduced in order to control the high number of DoFs, reaCog still relies heavily on the cooperation of individual parts, i.e., the combination of couplings between body, environment, the internal body model, and the controller itself. In addition, this internal model of its own body is used for planning ahead. Such a network, following Metzinger (2006, 2014) represents a system that is characterized by second-order embodiment.

As shown in figure 2, the procedures forming the decentralized controller are basically arranged in parallel, i.e., each procedure obtains its own sensory input and provides a specific

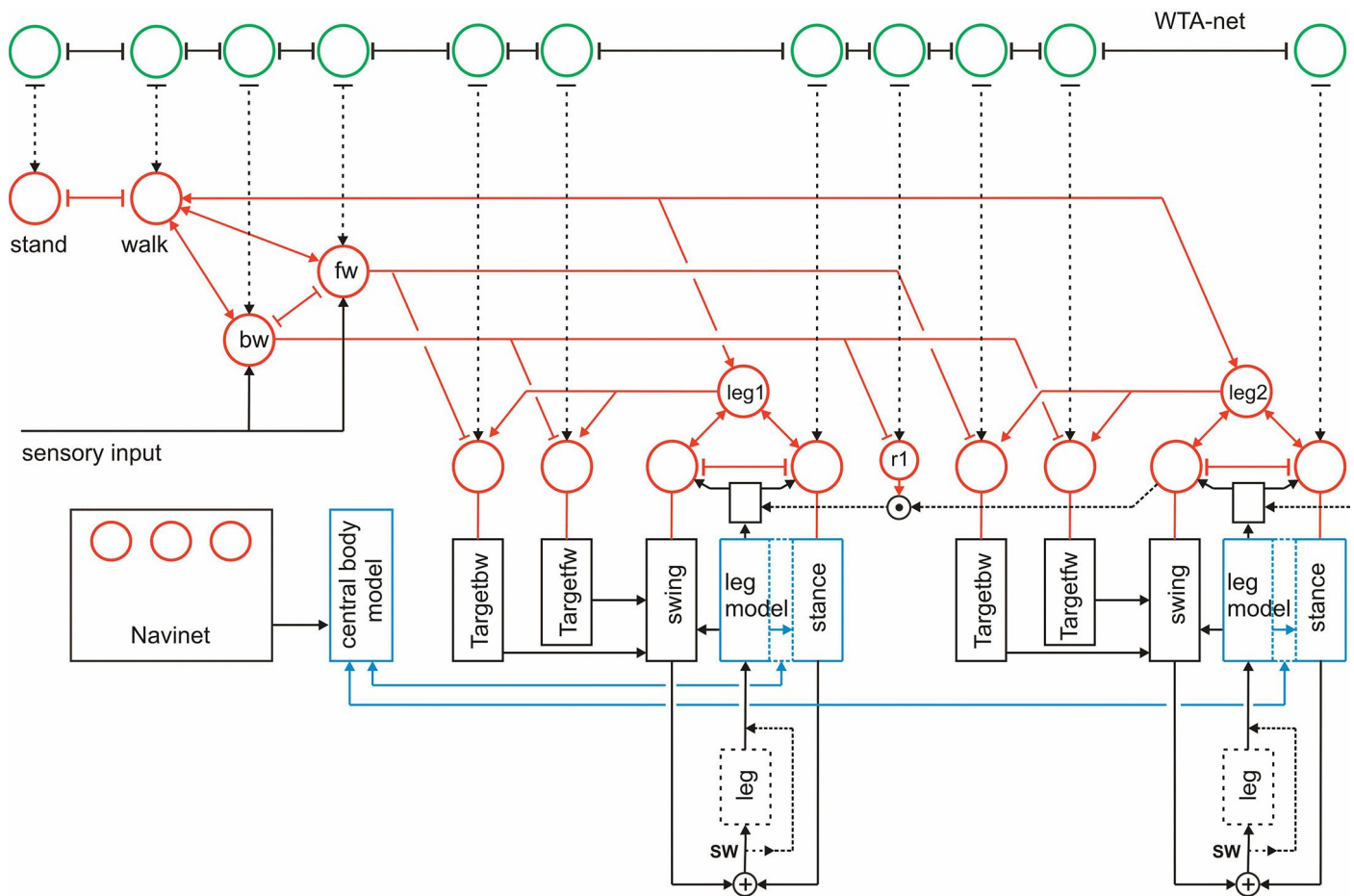


Figure 6: The controller of the reactive system as depicted in figure 2 expanded by a WTA-net (green units, not all connections are shown). Each WTA unit shows a bi-directional connection to a unit of the motivation unit network. This architecture provides the basis of reaCog, as explained in the main text. (for further explanations see figure 2).

motor output. But procedures can also receive input from other procedures and can provide output directly to other procedures. This relatively flat, heterarchical structure is also applied by the Word-nets and in perceptual memory (Cruse & Schilling 2010a).

ReaCog automatically selects actions on the lower reactive level. Several of these procedures can be performed in parallel. On the cognitive level, decisions about which action to choose are not based solely on sensory input, but are chosen depending on the imagined action, since there is a stochastic effect due to noise in the attention controller. The decision is afterwards tested by internal simulation before it is applied to the real system, and only after successful execution is the proposed behavior stored in long-term memory. Therefore, this decision process can be envisioned as a Darwinian type of selection that begins from stochastic

“mutations” that are then tested for “fitness” and selected based on this fitness. Thus, reaCog is a minimally cognitive system in the sense of the definition given by McFarland & Bösner (1993).

After we have defined the control network quantitatively, we can use reaCog to analyze emergent properties, which haven’t been implemented explicitly. As an example we have already considered a term like “tripod gait” that is sensible on a behavioral level in order to describe the emergent overall behavior of the walker. But on the control level there is no explicit tripod gait controller in reaCog (Schilling et al. 2008; Schilling et al. 2013a). The local influences coupling neighboring legs are responsible for overall coordinated walking behavior (different from many other hexapod controllers), and different gaits can emerge just by choosing different velocities. Therefore, appar-

ent “gaits” or the observation that “cognitive maps” are required can be seen an emergent property of such a network.

In the following, we will turn to concepts that are usually applied in fields different from computer science or behavioral biology, like psychology and philosophy of mind. Choosing another level of description can help us gain a better understanding of the system on a more abstract level. In addition, this approach can lead to more operational definitions for concepts used in other disciplines. This is based on the assumption that many of the above-mentioned phenomena emerge ([Vision 2011](#)) and that they can be used as concepts only on a higher, more abstract level.

For some authors, consciousness is thought to be restricted to human beings. In contrast, other authors share the opinion that there are degrees of consciousness and that consciousness does occur, to a smaller degree, in lower-level animals ([Dennett 1991](#)). Showing that quite small and simplistic networks can allow for interesting cognitive properties ([Chittka & Niven 2009](#); [Menzel et al. 2007](#)) supports such a view, as it provides a plausible evolutionary explanation for consciousness (or better degrees of consciousness). Agreeing with this basic assumption, we want to analyze to what extent our simple control network fulfils certain aspects of consciousness or emotions, even though we did not intend to realize this in our system in the beginning. The graded emergence of such high-level concepts would offer an evolutionary account and might allow us to address questions on the function, e.g., of consciousness, and explain how it relates to the control of behavior.

7 Phenomenality

Before concentrating on specific phenomena, such as emotions or consciousness, we would like to address a more fundamental aspect that appears to be relevant for all higher-level phenomena, namely the occurrence of subjective experience.

An example of subjective experience is pain. Even though it might be possible for us to closely attend to all neuronal activities of a hu-

man test subject while stimulating that person’s skin with a needle, the observed data would be different from the experienced pain, which is only felt by that person. Nobody other than that person can feel the pain. This form of experiencing an internal perspective is therefore only accessible to us through self-observation. Intuitively, other systems—like non-living things or simple machines—lack such an internal perspective. But in many cases, like for animals, it is hard to determine whether they have subjective experience or are merely reflexive machines that do not possess an internal perspective.

This problem is also visible when we consider a human brain, in the contrasting states of being awake or asleep, for example. While in (dreamless) sleep or under anesthesia the same neuronal systems as in a wakeful state may be active, subjective experience is assumed not to be present. And even in a normal wakeful state, we are not aware of all the contents of the different neuronal activities that take place in our brain. Therefore, only a specific type of neuronal activity seems to be accompanied by subjective experience.

There is only indirect evidence on the conditions required for subjective experience. [Libet et al. \(1964\)](#) performed an early experiment, where the cortex of a human subject was directly stimulated, electronically. Only for stimuli longer than 500 ms did the subjects report a subjective experience. Bloch’s law ([Bloch 1885](#)) formulates this connection more generally. The subjectively-experienced strength of a stimulus depends on the mathematical product of stimulus duration and stimulus intensity. In other words, a stimulus is only experienced subjectively when the temporally-integrated stimulus intensity surpasses a given threshold.

More recent experiments have studied the concurrent activation of different procedures that compete for becoming subjectively experienced. A basic experiment has been performed by [Fehrer & Raab \(1962\)](#), and has been followed by detailed later studies ([Neumann & Klotz 1994](#)). First, participants learned to press a button whenever a square was shown on a screen, but not when two squares were shown in a position on the screen flanking the first

square. After the learning period was over, in the experiment the single square was presented for only a short period (about 30 ms), which was then followed by a longer presentation of the two squares. The participants did not report having seen the single square, but reported only having seen the two squares. Nonetheless, they pressed the button. This result shows, first, that the first procedure A (“stimulus single square—motor response”), can be executed without being accompanied by subjective experience of stimulus stimA, the single square. Second, procedure B (“stimulus double squares—no motor response”) appears to influence how the first procedure is experienced, i.e., this procedure inhibits the subjective experience of stimulus stimA. Therefore, stimulus stimA is not subjectively experienced (the “masking” effect), but nonetheless triggers the motor reaction.

This situation can be interpreted in the following way (Figure 8, left). On the input side, each procedure shows temporal dynamics that are similar to that of a low-pass filter (LPF) (see footnote on page 2) followed by an integrator (IntA, IntB).⁶ Stimulation of one procedure inhibits the representation of the other procedure for some limited time (figure 8, Δt). In addition, both integrators are coupled via mutual inhibition (in figure 8 depicted by separate units). In the masking experiment, the first stimulus (stimA) does not inhibit the second procedure (B), because the latter is not yet stimulated, as long as stimulus stimA is active. In contrast, when the second stimulus, stimB, is given, the representation of procedure A may be suppressed. The representation of the input given by units IntA and IntB activate the corresponding motivation units (MU) of the procedures, MUA and MUB, respectively. This could be explained if we assume two different thresholds. First, the motor command of a procedure can be elicited when a small threshold (thr1, figure 8) is reached. But, a second, larger threshold (thr2, figure 8) must be reached in order to have subjective experience. Then, in our paradigm, procedure A, which was activ-

ated first, may reach the level of thr1, which is sufficient to activate the motor output, but not thr2. Only the second procedure, B, has enough time to reach the state of subjective experience (thr2, figure 8, right), which allows the double square (stimB) to become subjectively experienced (however this comes about). The model therefore suffices to explain the basic properties characterizing the backward-masking experiment. As has been shown by Cruse & Schilling (2014), the structure depicted in figure 8 can also deal with a forward-masking paradigm, the so-called attentional blink effect (Schneider 2013). To further describe another experiment, showing the so called psychological refractory period (PRP) paradigm (e.g., Zylberberg et al. 2011), the motivation units (MUA, MUB) of procedure A and procedure B are connected in such a way as to inhibit each other. In other words, the motivation units of these procedures form a WTA network. In addition, each procedure inhibits its own motivation unit after its action has been completed.

From these observations we conclude that there are specific neuronal states that require time to be developed. While eliciting an output signal (like a motor command) is the basic function of the system, this can happen without accompanying subjective experiences. Only some procedures may give rise to such phenomenal experience and might, in addition, trigger subsequent functions in the neural system. For example, this procedure may be able to access more neuronal sources and perhaps allow faster storing of new information (e.g., for one-shot learning). In addition to such functional properties the network can endorse the (mental) property of showing subjective experience, i.e., entering the phenomenal state.

The experimental findings mentioned above support a non-dualist, or monist, view, which means that there are no separate domains (or “substances”), such as the mental and the physical domain, in the sense that there are causal influences from one domain to the other one as postulated by substance dualism. Rather, the impression of there being two “domains”—often characterized as being separated by an ex-

⁶ An integrator performs a mathematical integration, i.e., it sums the input over time.

planatory gap (Levine 1983)—, results from using different levels of descriptions.⁷

An explanation of the necessary and sufficient conditions of neural networks that allow for subjective experience would be extremely interesting. Even though there currently exist only early insights or mere speculations, there has been a lot of progress during the past few years (review Schier 2009; Dehaene & Changeux 2011). The continuation of these research projects will hopefully yield a more detailed understanding. Using combinations of neurophysiological and behavioral studies may lead a better understanding of the physiological properties and functions of this state. It is, however, generally assumed that even if we knew the physical details at some future time, we would not understand why this state, which is characterized by physical properties, is accompanied by phenomenal experience. Here we propose another view. We assume that this problem will be “solved” such that the question concerning the explanatory gap will simply disappear, as has happened in the case of explaining the occurrence of life. Concerning the latter, there was an intensive debate between Vitalists and Mechanists at the beginning of the last century on how non-living matter could be transformed into living matter. The Vitalists argued that a special, unknown force, termed *vis vitalis*, was required. After many decades of intensive research, we are in a position where an internal model is available, which represents the observation that a specific collection and arrangement of molecules is endowed with the property of living. This and similar cases may be generalized as the following rule: If we have enough information, such that we can develop an internal model of the phenomena under examination, and if it is sufficiently detailed to allow the prediction of the properties of the system, we have the impression of having understood the system. In the case of life, indeed we do not need a *vis vitalis* any longer, but consider liveliness an emergent property. Correspondingly, we propose that if

we knew the functional details and conditions that lead to matter having subjective experience well enough, so that the appearance of subjective experience can be predicted, we would have the impression of having understood the problem. Therefore, we assume that the question of the explanatory gap will disappear at some point, as was the case in the example of life.

Adopting a monist view allows us to concentrate on the functional aspects when trying to compare systems endowed with the phenomenality, i.e., human beings, with animals or artificial systems. According to this view, phenomenality is considered a property that is directly connected with specific functions of the network. This means that mental phenomena that are characterized by phenomenal properties—as are, for example, attention, intention, volition, emotion, and consciousness—can be examined by concentrating on the aspect of information processing (Neisser 1967).

To avoid possible misunderstandings, we want to stress that we do not mean that the phenomenal aspect does not have any function in the sense that the system would work in the same way if there was no such phenomenal properties. Since, according to our view, the phenomenality necessarily arises with such a system, a version of such a system showing exactly the same functions but not having the phenomenal aspect would not be possible. A change in the phenomenal properties of a system has to be accompanied by a change in its functional properties. Functional and phenomenal aspects are two sides of one coin. However, remaining on the functional side makes the discussion much easier.

To summarize, the content of any memory element may be subjectively experienced (or available to conscious awareness) if (1) the (unknown) neuronal structures that allow for the neural dynamics required for the phenomenal aspect to occur are given, and (2) the strength and duration of the activation of the memory element is large enough, provided the element is not inhibited by competing elements.

The question of how any system can possibly have subjective experience was famously called the “hard problem” by Chalmers (1997).

⁷ There are various views adopting a monist approach, that differ in detail (epiphenomenalism, emergentism, property dualism and their many derivatives, see Vision 2011). We will not enter into this discussion here.

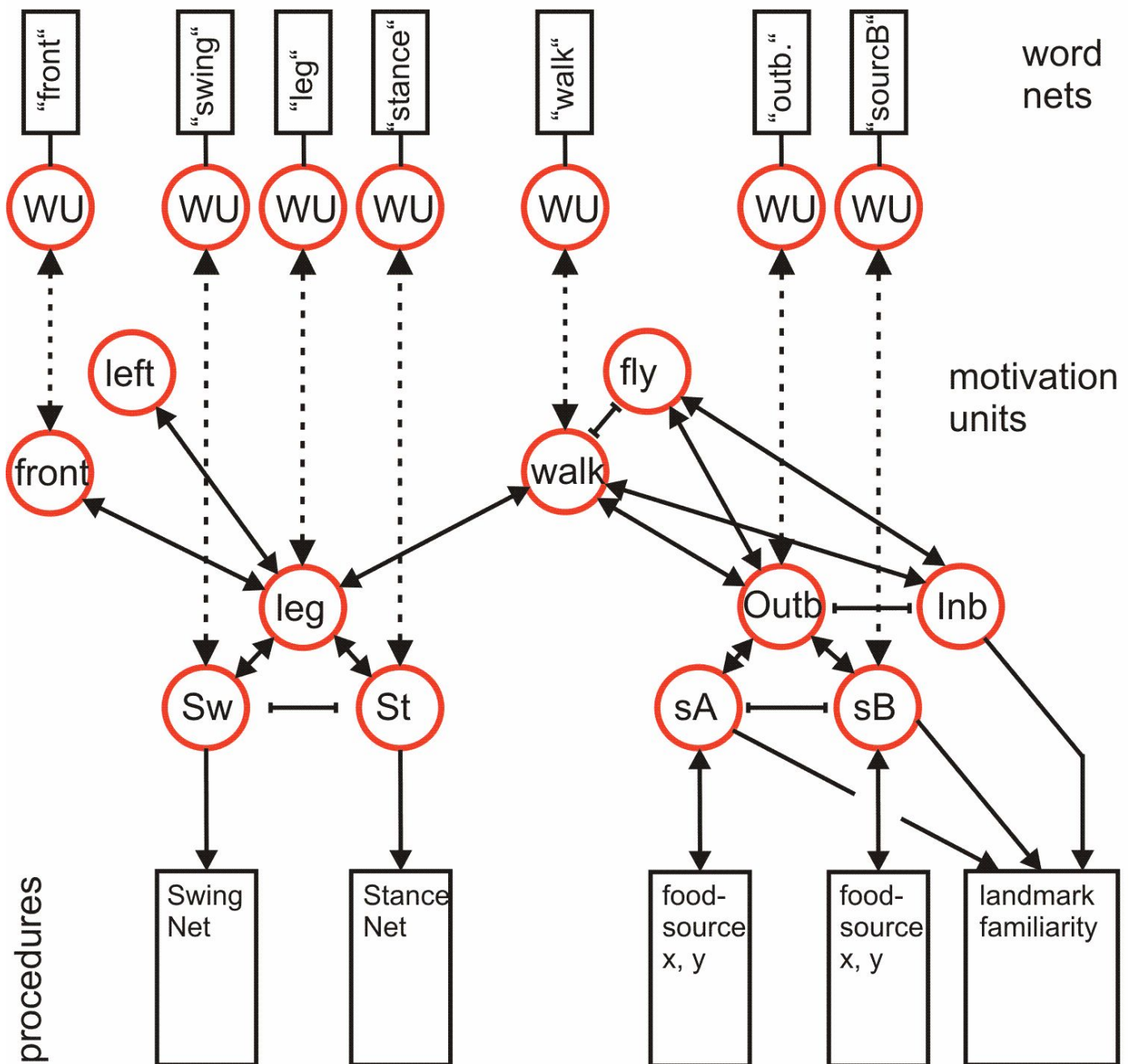


Figure 7: The reactive network expanded by a layer containing procedures that represent words (Word-net, upper row). The motivation unit of a Word-net (WU) is bi-directionally connected (dashed double-headed arrows) with the corresponding motivation unit of the reactive system containing procedural elements of Walknet (left, see figure 2) and of Navinet (right, see figure 4). The word stored in a Word-net is indicated as (“ ... ”). Not all of these motivation units have to be connected with a Word-net.

Adopting a monist view, we can avoid this question and leave it open, as we are interested in understanding the functional aspects of consciousness (on the ethical implications of an artificial system having subjective experience implemented in appropriate neural dynamics see Metzinger 2009, 2013). Regarding what kind of

dynamics could be thought of, it has been speculated that subjective experience might occur in a recurrent neural network that is equipped with attractor properties. Following this hypothesis, subjective experience would occur if such a network approached its attractor state (Cruse 2003). This assumption would mean that any

system showing an attractor might be endowed with the phenomenon of subjective experience. It may, however, not have all the other properties characterizing consciousness. On the other hand, there might be systems in which the functional aspects currently attributed to consciousness are fulfilled, but where there is no subjective experience present. This case would imply that our list representing the functions of consciousness as given in section 10 below is not yet complete.

In the following two sections we shall briefly treat two phenomena—emotions and consciousness—and discuss how they might be related to the minimally-cognitive system as represented by reaCog.

8 Emotions

Most authors generally agree that emotions are accompanied by subjective experience and that they have the function of helping the subject respond adaptively to environmental pressures. So there is the phenomenal aspect of emotions as well as a functional aspect. As we have already treated the phenomenal aspect above, here we will put aside this aspect, i.e., how it feels to be happy, sad, etc., and concentrate on the functional aspect of emotions.

Even though several authors assume or even demand that emotions are already present in simple reactive systems, and that they are necessary for a cognitive system (Valdez & Mehrabian 1994), in our above description of the properties of the network reaCog, any emotional aspects have not been taken into account. We did not require the term “emotions” to explain our approach, nor have we built in any kind of explicit emotional system. However, we will argue that there are emerging properties that are comparable to what is usually ascribed to properties of emotional systems. In the following, we want to focus on which parts in our system take this role and how the functions of these parts can be described and related to attributes of emotional systems.

The attempt to relate the properties of our network with the concept of emotions appears not very promising at first sight, because

a series of interrelated conceptual terms such as emotions, attitudes, motivations, sentiments, moods, drives, and feelings can be found in the literature, and are defined in different but partly overlapping ways by different authors (Pérez et al. 2012). The reason for this disagreement might be that there are indeed no clearly separable mechanisms underlying these phenomena but rather we are dealing with a holistic system, which makes separation into clear-cut concepts difficult, if not impossible. As mentioned, the problem of being confronted with heterarchical structures appeared when looking at the reactive level (and reappeared later when dealing with perceptual memory), which led us to the neutral term “motivation unit” for all “levels” of the heterarchy formed by the motivation unit network. To simplify matters, we will only deal with the term emotions in the following.

What might be possible functions of emotions? As follows from the examples of overlapping conceptual approaches found in the literature and mentioned below, emotions are attributed to various functions characterized by different levels of complexity. These range from enabling the agent to select sensory input (e.g., tunnel vision, Pérez et al. 2012) and activate different procedures, or, at a higher level, to select between different behavioral demands (e.g., hunger – thirst, flight – fight, Parisi & Petrosino 2010) up to more abstract states such as suffering from sadness or being in a state of happiness and controlling the corresponding behaviors (e.g., Ekman 1999). The lower-level decisions are well covered by our motivation unit network, and form a heterarchical system showing attractor states (e.g., swing – stance, Inbound – Outbound). These states allow for selection of sensory input and/or motor procedures that are stimulated by sensory input to specific motivation units. In the following, we therefore focus on higher-level states, such as, for example, emotions, as listed by Ekman (1999).

In general, and as discussed below, one can distinguish between prototypical approaches and reductionist approaches—the latter simplifying emotions down to just a few basic dimen-

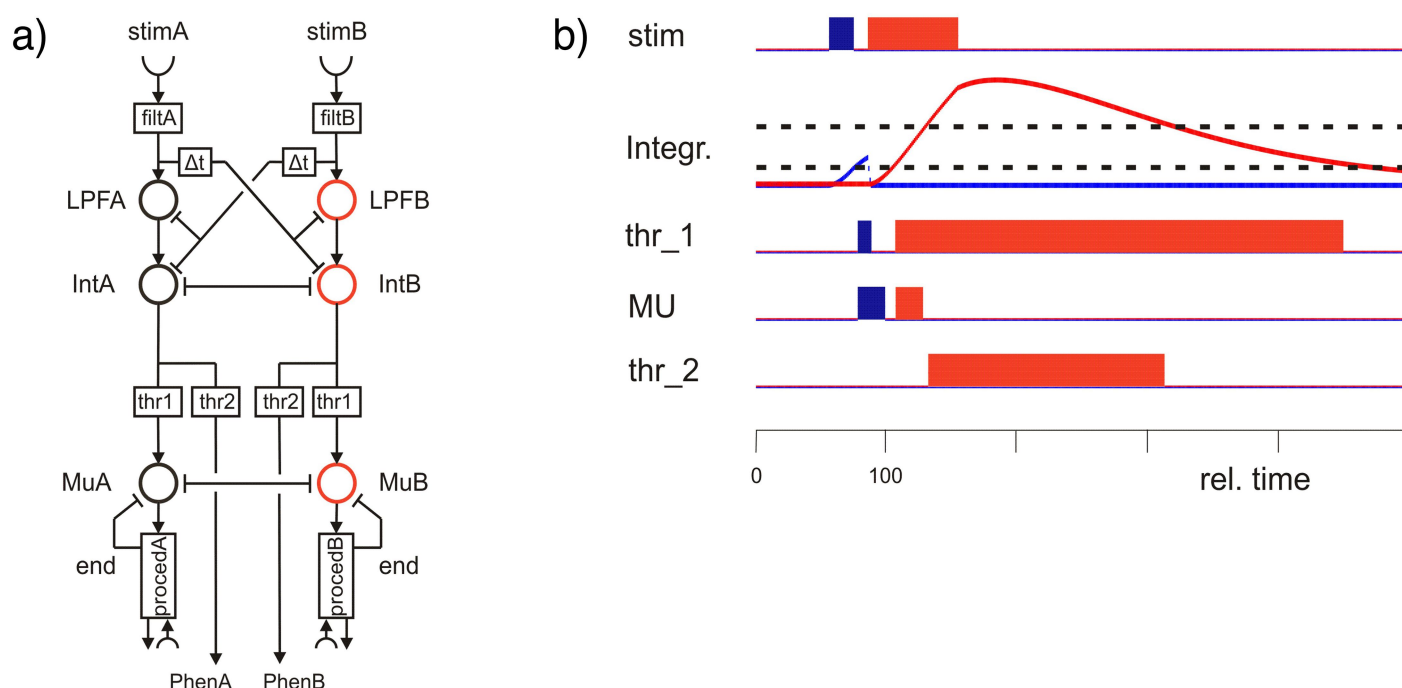


Figure 8: (a) A hypothetical network that is capable of dealing with some dual task experiments, for example the backward masking experiment. Stimulation of one of the procedures, A or B, activates a low-pass filter (LPFA, LPFB) followed by an integrator (IntA, IntB) and inhibits the corresponding units of the other procedure for a limited time (Δt). The integrators are coupled via mutual inhibition. After activation of one of the integrator units has reached threshold thr1 (lower dashed line), the corresponding motor motivation unit (MuA or MuB), coupled via mutual inhibition, is activated, which drives the behavior. If threshold thr2 (upper dashed line) is reached, the stimulus can be phenomenally experienced. A feedback from the procedure can provide an “end” signal to inhibit its own motivation unit. (b) Temporal development of the activation of some units (procedure A, blue, procedure B, red). Abscissa is relative time. If stimB follows briefly after stimA, the unit IntA may reach its motor threshold thr1, but not the threshold thr2 for eliciting the phenomenal experience. In contrast, stimB elicits both the motor output and the phenomenal experience that corresponds to the backward masking effect (for details see [Cruse & Schilling 2014](#)).

sions. In current research, both views appear to be justified as they both try to describe the phenomena observed, though at different levels of description.

Following the first approach, research tries to trace emotions back to a set of basic emotions, the combination of which can explain further derived emotions. This approach has been advocated by [Plutchik \(1980\)](#). A problem with such an approach is how to draw borders between emotions and what counts as a basic emotion. [Ekman \(1999\)](#) proposed a list of characteristics of similarity between emotions and came up with a set of fifteen basic emotions. Later on, based on their relation to facial expressions, he reduced this number to six. This set, which is now widely used as the basic set of emotions in many different contexts, consists of

happiness, anger, disgust, sadness, fear, and surprise. As an example, let us consider happiness. Happiness is elicited when we are in a state of having had or expecting positive situations. The behavioral effect of happiness might be characterized as being open to new ideas, perhaps not being too critical and open to performing new, unconventional behaviors. How might such a phenomenon be represented in reaCog? First of all, a neuronal state of the motivation network would correspond to a specific emotion. Such a network state is usually triggered by some sensory stimulus eliciting an emotion. This stimulus activates specific, basically innate, networks which, when active, influence the system and put it into the respective emotional state. Such a network—which could, in the most reduced case, consist of just one neuronal unit—has not

been introduced in reaCog, but if assumed as given, it may modulate meta-control parameters such as, for example, noise levels, thresholds, or learning rates (Doya 2000, 2002). To stick to our example of “happiness”, activation of such a network, which represents stimulus situations considered to elicit this state may, within the Spreading Activation Layer, lead to a faster diffusion process, perhaps supported by stronger noise amplitude. Such a broadening of the attention range as a consequence of positive affects has been reported by Dreisbach & Goschke (2004). In addition, or as an alternative, the threshold for the problem detectors that we mentioned in section 4 might be increased. As a consequence, the system would take more risks. All these changes would lead to an increase in “creativity”, i.e., the ability to find new ideas for possible solutions. Corresponding structures might be found in the other basic emotions listed by Ekman.

In the second group of approaches to characterizing the emotions, emotions are described through a set of dimensions that represent the emotional state. We will briefly sketch this seemingly alternative reductionist approach and will again draw parallels with reaCog. The connection to reaCog is made on a different level and is therefore not logically exclusive with respect to the former. Wundt (1863) was quite opposed to the idea of breaking down emotions into a set of basic emotions that serve as prototypes, mainly because he assumed that a set of emotions is better described by a continuum than by separable categories. This follows his idea of describing emotions through principal components leading to dimensional systems, like the pleasure-arousal-dominance (PAD) framework (Mehrabian 1996). In the PAD framework, three dimensions span the space of the emotions. The first describes the state pleasure–displeasure and corresponds to the affective state (excited – relaxed). Arousal, as the second dimension, represents the level of mental alertness and physical activity (tense – sleepy). The third axis describes the level of dominance–submissiveness, i.e., the feeling of being in control. The three factors of the PAD framework have successfully been employed as semantic differen-

tial factors to describe emotional states in different contexts, e.g., for describing postures, facial expressions, gestures, and vocal expression. The three dimensions appear to be sufficient as they capture large parts of the variance (Mehrabian 1996). Mehrabian has related the three traits—pleasure, arousal and dominance—to specific cognitive characteristics. First, pleasure–displeasure, according to Mehrabian, deals with the fulfillment of expectations. Fulfillment of an expectation (or not) occurs when, during a problematic situation, planning ahead is activated and after some time and searching a solution is found (or not)—a state that can be found in reaCog, too. But fulfillment of expectation might also occur at lower levels, when, for example, a simple procedure such as Swingnet is equipped with a target value and this goal is either reached or not. The error signal might then be used as a measure for fulfillment of expectation. For example, it might be used as problem detector in the case mentioned earlier, when a subject tries to lift a leg off the ground, but due to an inconvenient load distribution, the body falls down and the leg remains in contact with the substrate. The arousability trait, as introduced by Mehrabian (1996), was meant to incorporate the process of “stimulus screening”. In short, “stimulus screening” is a process of attentional focusing. Such a process of focusing attention occurs in our system, too, as, on the one end of the spectrum, the system broadly attends to all environmental influences as perceived through its sensors, and this is characterized as its being in the “reactive state”. At the other extreme, when a specific problematic situation occurs, it is necessary to focus attention and to guide the search for a solution towards specific modalities, parts of the body, etc. But even on the reactive level, attention selection can be observed, as we mentioned earlier. Finally, the dominance trait (“generalized expectations of control” Mehrabian 1996) concerns the extent to which the agent takes over in the actual situation and is not only responding and reacting, which agrees with the main thesis of our approach, namely that it is possible to switch between the reactive mode and the cognitive mode. Similarly, Russell &

Norvig (2003) have required an autonomous agent to be able to both react to known situations and to be in control of the situation itself (or as Russell and Norvig call it: being proactive).

Our approach, as we have mentioned, does not aim to build specific emotional properties, but tries to build a functional autonomous system and then to look at the aspects of emotional properties that might be found in the network or gained after some further functional expansion of the network. We have listed examples from different levels of description in psychology and point to related properties in our network. We are not arguing that reaCog has emotions (we are in any case agnostic with respect to the subjective aspect). Rather, we claim that by taking a network like reaCog as a scaffold, different conceptualizations of the functional aspects of emotions can be mapped onto such a quantitatively defined system and thus be considered emergent properties.

It might be added here that recent studies support the idea that emotion-like states do indeed occur in brains which are by far less complex than mammalian brains. Yang et al. (2014) could show that the concept of “learned uncontrollability”, generally considered as an animal model for depression as observed in humans, can be found in *Drosophila*, too. For vertebrates, it is known that stress induces the state of fear or of anxiety, the latter being considered as a second order emotion. Fossat et al. (2014) could show that a crayfish treated by stressors (i) avoids illuminated parts of the environment and (b) shows an increased level of serotonin in the brain, as can be observed in vertebrates. As in vertebrates, the state of anxiety could be relieved by application of anxiolytic drugs. Both results have been interpreted such that the ability to adopt emotional states must have been evolved before the separation of the arthropods and vertebrates.

9 Attention, volition and intention

In the following section we want to turn to attention, intention, and volition. To what degree can those properties be attributed to our sys-

tem? We start from the definitions of attention provided by Desimone & Duncan (1995), of intention from Pacherie (2006) and Goschke (2013), and of volition from Goschke (2013).

Attention is the ongoing selection process in perception. It can be driven bottom-up, i.e., by sensory influences, or it can be controlled by top-down influences (Desimone & Duncan 1995). Top-down driving of attention depends on the internal or emotional state and might depend on familiarity with the stimulus.

We can indeed find properties corresponding to attention in reaCog. The motivation network is constituted of local clusters of units that always compete on this local level and form in this way coalitions of units and small subclusters. As an example, we introduced the selection of procedures at the leg level. Either a swing or a stance motivation unit can be active and inhibits the other one. These two units compete for control of behavior. Sensory units can influence this competition. For example, an incoming ground-contact signal ends a swing movement and initiates stance activation. After activating the “Stance” unit only sensory input relevant to stance can be perceived by the system, but not inputs relevant to swing. Therefore, this case corresponds to bottom-up attention control.

Such competition can also be found on a global level, on which different behaviors can be chosen. The activation of these higher-level elements influences the lower level. This activation provides a context for the lower level, which guides the selection process on that level and decides which sensory inputs might be relevant. Thereby, more global clusters control the attention on the lower levels in a top-down fashion. Corresponding examples can be found in Navinet, which we mentioned earlier. Only visual signals concerning landmarks that belong to the current active context are considered and switching between contexts only becomes possible after the food source has been depleted and found empty.

The cognitive expansion of reaCog represents another case of top-down influence. This system comes up with new behaviors and probes them via internal simulation. As men-

tioned, there is a specific WTA layer that mirrors the arrangement of the lower motor control layer (figure 6, green units). This part of the controller can be called an “attention controller”, as the explicit function of this layer is to narrow down the search for suitable behavior and to actively select a single one. We call this selection a cognitive decision, as the system is supposed to select a behavior that would not normally be triggered through the given context. In this way the system represents a special type of top-down attention. The focusing mechanism may correspond to what sometimes has been termed “spot light” (Baars & Franklin 2007 p. 955). Overall, we can therefore observe three different types of attentional influences in reaCog.

Volition is an umbrella term denoting mechanisms allowing for voluntary actions. The latter are “actions that are not fully determined by the immediate stimulus situation but depend on mental representations of intended goals and anticipated effects” (Goschke 2013). For an outside observer, voluntary actions cannot be predicted. As mentioned above, it is crucial for the cognitive expansion that it can select behaviors that are not triggered by the current situation. The system has to invent new behaviors. Even though the consequences of these behaviors are predicted, from the outside the finally chosen behavior is not predictable, as this invention and selection of new behaviors is stochastic to some extent. The application of internal simulation only guarantees that the proposed behavior will lead to a solution, but it does not give away which behavior will be chosen. To the contrary, the search space of possible solutions can easily become very large and has to be restricted. Such restrictions help to span a tractable space of possible solutions. In our example, reaCog looks first for solutions in the morphological neighborhood, i.e., it tries to use the neighboring legs to help find a solution for a locally-given problem. There are still many possible behaviors that must be tested in a somewhat random order. The system will end up with one that has been anticipated as a solution in internal simulation, but this solution is not selected through sensory inputs or the current con-

text as such. Therefore, volition may be attributed to a system like reaCog.

Does an agent controlled by reaCog show intentions? Intentions are present when the controlled action is goal-directed. We are following Pacherie (2006), who proposes a differentiation of three types of intentions (based on Bratman’s (1987) original differentiation into two such types). Pacherie distinguishes future-directed as well as present-directed intentions and introduces motor-intentions as a third type. Present-directed intentions are considered to be under “conscious” (or “rational”) control. In contrast, motor intentions are related to lower-level function (Pacherie 2006). Defining for these types of intention is that they provide guidance for the function on the respective level. In reaCog, motor-intentions are realized by the fact that, on the reactive-control level, behaviors can be selected based on the context. Present-directed intentions can be found on the level of cognitive decision. Future-directed intentions are not treated by reaCog, because its architecture in the current version only allows for dealing with problems that occur in the context of current walking behavior. However, an expansion of reaCog that would include planning ahead using Navinet as a substrate would include future-directed intentions, too.

Goschke (2013) defines intentions as “causal preconditions explaining why a particular stimulus triggers a particular action (rather than a different action)” (Goschke 2013, p. 415). In other words, “intentions can be said to shape the “attractor landscape” of an agent’s behavioral state space” (Kugler et al. 1990, ref. from Goschke 2013, p. 415). In reaCog, such an attractor landscape is described by the motivation unit network. As explained in the preceding paragraph on attention, the activation of a context guides, in a top-down fashion, both the selection of a suitable behavior as well as which sensory inputs the system should attend to. The lower-level activation and incoming sensory inputs influence, on the one hand, the adaptive execution of the behavior as such. On the other hand, the sensory input can inform the higher level in bottom-up fashion and might indirectly trigger changes on this higher-level, too. The ac-

tivation on the higher level will, however, be in general more stable on a temporal scale and will reflect a specific context as well as relate to specific goals. For example, in the case of Navinet, there are different possible goals, such as food sources or the nest, which are represented in the higher-level network. Selecting one of these as a goal will guide the overall function of the system, as its behavior is directed towards approaching that location, while the sensory system will attend only to the specific (expected) sensory stimuli. Therefore, reaCog can be assumed to show goal-directed behavior and intentions.

10 Consciousness

In this section we would like to discuss to what extent properties of consciousness might be found in our system. Even though we start from a common notion of how consciousness can be viewed as consisting of separable domains, we are well aware that this approach is not the only or ultimate solution for approaching this question. But such a differentiation appears well-suited for our bottom-up approach.

Overall, many authors contribute to the view in which consciousness is broken down into a set of properties. We start from a review by Cleeremans (2005), who gives a good overview on the diverse philosophical views on consciousness and tries to integrate them into one framework. While there is disagreement in general and also on the details (see also Vision 2011), Cleeremans interestingly finds a common denominator between the different opinions that characterize possible computational correlates of consciousness. He introduced a differentiation of consciousness into three domains: phenomenal consciousness, access consciousness, and metacognition (or in other contexts referred to as reflexive consciousness). There is disagreement on the phenomenal aspect, as it is seen by one group of philosophers to be an independent domain. In contrast, there is also a view in which phenomenality cannot be separated from metacognition and access consciousness, but must be seen in relation to those (see review Cleeremans 2005).

We have argued in section 7, that the phenomenal aspect as such, i.e., the property of some neuronal structures that are equipped with subjective experience, has *per se* no function, but is, nonetheless, not separable from the functional properties. Therefore, we see the phenomenal aspect not as a separate type of consciousness, but as a property of both access consciousness and metacognition. This view has convincingly been supported by Kouider et al. (2010) as well as, in a recent review, by Cohen & Dennett (2011). Therefore, we will compare properties of reaCog with current definitions found in the literature concerning the phenomena of access consciousness and metacognition, abstracting from the phenomenal aspect.

While other philosophers require metacognition or reflexive consciousness in a system in order to attribute consciousness (see for example Rosenthal 2002 or Lau & Rosenthal 2011 for a recent review defending this view), we do not want and cannot get into this discussion as it is not our goal to review the different types of taxonomies. We basically follow one valid and common perspective, as presented by Cleeremans, and apply it to our system in order to analyze functions of our system that can match the different phenomena described. We do not aim with this approach to give a rigorous definition of consciousness (which does not seem suitable at this point, see also Holland & Goodman 2003). Instead, applying our approach, we aim to provide insight into specific functions of our system that are connected to the phenomena discussed.

10.1 Access consciousness

In this section we want to focus on the aspects of access consciousness that can be found in reaCog. Following Cleeremans, access consciousness of a system is defined by the ability to plan ahead, to guide actions, and to reason, as well as to report verbally on the content of internal representations. In contrast, non-conscious representations cannot be used this way. Selecting behaviors, planning ahead, and guiding actions are the central tasks of reaCog (see section 4, Planning ahead).

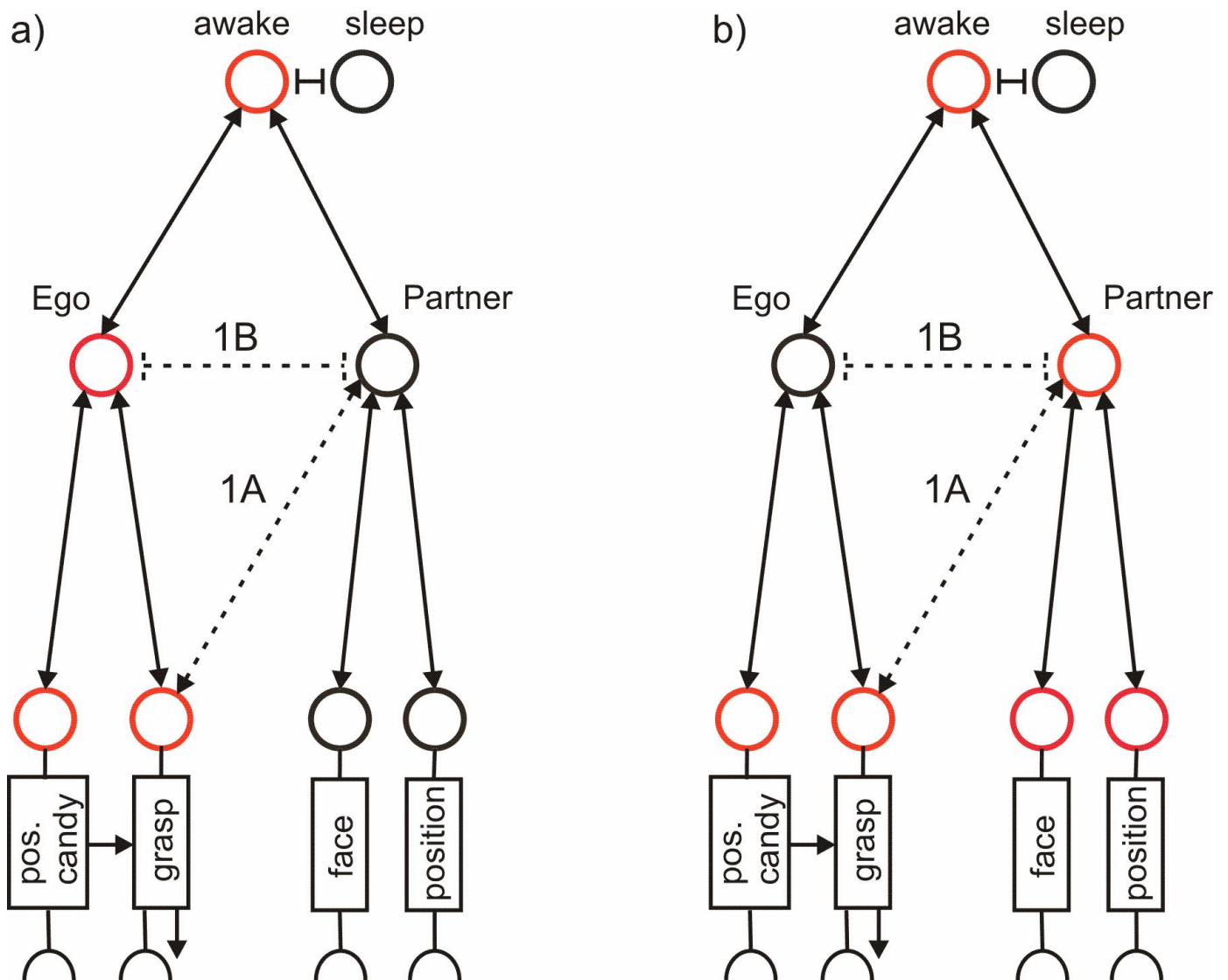


Figure 9: A possible expansion of reaCog. Without the connections 1A and 1B the network enables the agent to represent its own actions, as is already possible for the network shown in figure 2, and figure 6. After introduction of connections 1A and 1B the network is also able to represent the actions of a partner using the now shared procedure “grasp”. (a) and (b) show two attractor states where active motivation units are depicted in red, whereas inactive motivation units are shown in black. Half circles indicate sensory input.

Being able to use internal representations for verbal report is currently not a part of reaCog. However, the internal representation of reaCog is already suited to allow for accessing internal representations (section 5 and figure 7). The simple solution proposed allows for communication using one-word sentences only, but provides a way, within the framework of reaCog, for the symbol-grounding problem to be addressed. Steels (2007; Steels & Belpaeme 2005) and Narayanan (1997) have already studied in detail how more complex sentences may be grounded in simple reactive systems. Thus,

there already exists work on similar systems that shows how the ability to report by using more complex language structures could be implemented in a reactive system. Therefore, at least in principle, this property could be realized in reaCog, too.

The last property describing access consciousness, symbolic reasoning, is not addressed by reaCog. In the symbolic domain, there are, however, many interesting approaches in the literature that might be connected to a system like reaCog after the symbolic level has been implemented.

Concerning related work, [Dehaene & Changeux \(2011\)](#) review relevant network models that are supposed to simulate consciousness, including their own approach, which is termed global neural workspace theory (GNW) (see also [Seth 2007](#) for a systematic summary). A comparison of reaCog with these approaches can be found in [Cruse & Schilling \(2013\)](#)). Here we will only refer to one important notion, “global availability” as used by several authors to represent a crucial property of access consciousness (e.g., [Dehaene & Changeux 2011](#); [Dehaene & Naccache 2001](#); [Baars & Franklin 2007](#); [Cleeremans 2005](#)). Global availability describes the notion that many representations of the system can potentially become conscious. These representations can be selected to solve a current problem (as described for reaCog) or could be selected in a task (see GNW).

Are the representations used in reaCog globally accessible? During execution of a form of behavior the reactive system simply reacts to sensory inputs. Single local modules of the procedural memory are activated by the context, for example, the walking behavior that can execute walking even in a cluttered environment. While the behavior is driven by sensory stimuli, it is not “cognitively attended” and runs automatically in response to direct interaction with the environment. In this case, the representations are not attended by cognitive expansion and are clearly not a part of access consciousness. But, importantly, this can change whenever a problem is detected and the reactive (automatic) system is not sufficient anymore. In such a case, the WTA-net of the attention controller is activated and has to select one of the elements of the procedural memory. During planning, these elements become accessible to the attention system ([Norman & Shallice 1986](#)). The WTA-net, which constitutes the essential part of the attention controller, projects directly back to the motivation units of the procedural memories (figure 6, dashed arrows) and thereby selects just one of the possible behaviors (due to the characteristics of a Winner-Take-All network). Therefore, all the procedural modules that could be activated by the attention con-

troller are “globally available” and form possible elements of access consciousness.

10.2 Further relations between reaCog and access consciousness

Another interesting property of reaCog and findings in psychology concern the relation between conscious and automatic procedures. It is well known that humans are able to learn a new behavior by consciously attending to that behavior. Over time, this can change and the execution of the behavior becomes more and more automatic, i.e., it is no longer necessary to be consciously aware of the exact execution of the behavior. A similar shift of attention can be found when reaCog is planning new behaviors. Triggered by the activation of a problem detector, reaCog has to shift its attention towards the new behavior during planning and the following execution of a behavior. As long as the problem-detector is still active, the reactive system is basically suspended (by switching off the loop through the body), and instead the planning system tries out new procedures that have to be attended to. After the successful execution the new solution can also be stored as a procedural memory and become part of the reactive system; it does not require cognitive attention anymore (the procedure how to store this information has not yet been implemented in reaCog). An advantage of this integration into the reactive system is that access to reactive procedures is faster than using the cognitive process, which agrees with the findings mentioned above.

There are other experimental findings highlighting the relation between conscious and non-conscious access to procedural elements. [Beilock et al. \(2002\)](#) found that athletes who have learned a behavior so that it can be performed automatically perform worse when they concentrate on the behavior compared to when performing the behavior while being distracted. In the attention controller of reaCog we can observe a similar phenomenon. If the attention controller is externally activated by a higher-level unit while the connected behavior is performed, this could possibly activate learning.

Such an influence would change the underlying neuronal module and could worsen the result. In contrast, without attention the behavior would be performed as it had been learned earlier.

ReaCog differs in an important aspect from the simulation studies conducted by Dehaene and colleagues, as well as from those conducted by Baars and colleagues. While the latter approaches aim to relate conscious functioning to individual brain areas or brain circuits, reaCog is not intended at all as a model of the human brain or any of its areas. Instead, it is envisioned as a reductionist approach that focuses only on function. From the bottom-up development of more and more higher-level function we offer a post-hoc discussion of the question of to what extent reaCog shows aspects of access consciousness. This approach seems particularly suitable for addressing access consciousness, as it turns out that there is no single identifiable part of reaCog that might be attributed the property of access consciousness. Instead, access consciousness appears to be an emergent property constituted by the complete system. Attention controller, procedural memory, and the connections between those two parts, as well as the internal model and the ability to use it in internal simulation, seem to be the required structures that allow access consciousness, or, in other words, together constitute the “neural workspace.” The dynamics of the neural workspace as defined by Dehaene & Naccache (2001) are given through the WTA-net. But, and this is an important difference, there no re-representation in this neural workspace is necessary. The already-present representations can be reused in novel contexts. The existing modules of procedural memory are recruited in the internal simulation when planning ahead. The only difference is that the body is decoupled from the control loop and instead the loop through the world is replaced by a loop using internal models and their predictions as feedback. Together, these representations form the global workspace (this notion of internal models has been termed “second-order embodiment,” c.f. Metzinger 2014).

Koch & Tsuchiya (2007) differentiate attention and consciousness, as both can be present individually and independently of each other. They conclude that different mechanisms are responsible for attention and consciousness. While such a differentiation is of course based on basic definitions, we can indeed identify different mechanisms related to these two phenomena, even though they seem to be related. In reaCog, attending to a specific stimulus is modelled as a specific activation of motivation units. Only if this activation is strong enough and/or active for enough time, can the procedure enter the phenomenal state (section 7, figure 8). Therefore, both attention and the phenomenal aspect of consciousness refer to different, but tightly coupled properties of our system.

10.3 Metacognition

Although in this article we use the term cognition in the strict sense as proposed by McFarland & Bösner (1993), when dealing with metacognition, this definition is no longer generally applicable. Therefore, in this section the term cognition is used in the usual, more qualitatively-defined way. We will describe how the motivation unit network could be expanded to allow our agent to be endowed with different aspects of metacognition. These expansions, however, have not yet been simulated by being implemented into the complete network.

Metacognition, or reflexive consciousness (sometimes called metarepresentation), the second essential domain of consciousness, according to Block (1995, 2001) and Cleeremans (2005), is characterized by Lau & Rosenthal (2011) as “cognition that is about another cognitive process as opposed to about objects in the world” (p. 365).

While the selection of procedures for control of behavior may occur on the reactive level or by application of access consciousness, metacognition in addition is able to exploit information concerning a subject’s own internal states. As a further property, a metacognitive system, when selecting behavior, can represent itself *as* selecting this behavior (“I make the decision”). Metzinger (2014) classifies this ability as third-

order embodiment, where the subject's own body is "explicitly represented as existing" (p. 274) and the "body as a whole" can turn "into an object of self-directed attention" (p. 275). Thus, metacognition is about monitoring internal states in order to exploit this knowledge for the control of behavior. According to [Cleeremans \(2005\)](#), metacognition may also be used for inferring knowledge about the internal states of other agents from observing their behavior and for communicating a subject's own states to others.

Let us first focus on the individual agent. What kind of information might be used by a metacognitive system? A typical case discussed in the literature concerns some quality measure of the procedure to be selected. During decision-making, a person, when relying on own knowledge, needs to be able to access his or her own internal state in order to estimate how sure he or she is about the specific piece of knowledge. [Cleeremans et al. \(2007\)](#) use as an illustrative example a system consisting of two artificial neural networks. While the first network learns an input-output mapping of the task, the other network, as a second-order network, learns to estimate a quality measure describing the performance of the first-order network. As the combination of the two networks does not only store information in the complete system, but also contains information about and for the system, the authors conclude that such a system already shows a limited form of metacognition. Such a network, using an additional second-order subnet, might be implemented in our system, too. For example, motivation units could be activated by confidence, or quality values estimated by such a second-order network. Such a situation can indeed be found in the network Navinet. Navinet is used for navigation control tasks and is inspired by work on navigation in ants. In this system, the salience of a stored stimulus guides memory retrieval ([Cruse & Wehner 2011](#); [Hoinville et al. 2012](#)). For instance, the decision to choose one of many different food sources is influenced by the internal representation of the learned food quality ([Hoinville et al. 2012](#)). As another example, the confidence value of a visual landmark that is to

be followed or not might depend on the salience of the visual stimulus, similar to the implementation of a Bayesian-like system. A different example is given by reaCog, which, by exploiting its internal body model, is capable of representing its own body for internal simulation as well as for control of behavior. Thus, at least some basic requirements for metacognition, such as being able to use own internal representations for the control of behavior, are fulfilled, if we, again, leave the phenomenal aspect aside. Below we will, in addition, briefly address the ability of the agent to represent itself.

How may metacognition be suited to support information transfer between different agents? We will not refer to communication using verbal or gestural symbols here. Instead, we want to start with the ability to identify oneself with another agent, or, in other words, to be able to "step into the shoes of the other." This faculty has been referred to as Theory of Mind (ToM). Central is the notion of being able to attribute mental states to other agents ([Premack & Woodruff 1978](#)). A classical example is the "Sally–Anne task". In this experiment, two subjects observe how a cover hides a piece of candy lying on a table. While one subject, Sally, is outside of the room, the other subject, Anne, is able to observe how the hidden candy is moved to a new location. After the change the candy lies underneath a white cover and not under the black cover, which it did to start with. The crucial test question is put to Anne: where does she think Sally will search for the candy? If Anne points to the white cover she only uses her own current beliefs about the situation, but does not apply a ToM, i.e., she does not take into account what Sally believes—since Sally has not observed the switch. But if Anne points to the original location, the black cover, she is assumed to have a Theory of Mind as she operates on a set of mental states that she ascribes to Sally.

ToM is crucial when an agent needs to capture not only physical objects, but in addition represent other agents. It becomes necessary to explicitly keep track of others' observations, plans, and intentions. Only such agents that can attribute mental states to other agents

can successfully predict their behavior. There are two common explanations to account for how ToM is realized. First, the so-called theory-theory (Carruthers 1996) assumes that there are dedicated, innate, or learned procedures that allow for prediction of internal states and therefore the behavior of others. We want to concentrate on the second main approach, namely simulation theory (Goldman 2005).

Central to simulation theory is the already introduced notion of an internal simulation. As a prerequisite an agent needs an internal model of him or herself. This model can be used (as explained) for planning ahead using internal simulation. But in the same way this model can also be recruited in order to represent another agent. Thereby, other agents may be mapped onto the own internal model that allows simulating the behavior of the other agent. This faculty would enable the agent to derive all sorts of conclusions based on its own representations, such as, for example, current goals or intentions.

In the case of reaCog, we envision an extension that allows mapping another agent onto the already existing internal model. Internal simulation could be used in this context, too. Therefore, the application of such an internal simulation of another agent could lead to an interpretation of the behavior of the other. However, the two theories mentioned do not necessarily exclude each other, as can be shown when regarding the properties of the cognitive expansion further. If the interpretation found via an internal simulation of another agent is new and succeeds in simulating its behavior, the result could be stored in the procedural memory in a similar way as described for reaCog, when coming up with a new solution to a given problem. In this way, a new procedure has been learned that allows for prediction of the behavior of the other agent. As such, application of simulation theory might in the end lead to results that are described as characterizing theory-theory. The faculty of applying a ToM is currently beyond the ability of reaCog as described above, which allows for an egocentric view only. In the following, we will, however, sketch a way in which such a network may be implemented

into the architecture of reaCog (for more details see Cruse & Schilling 2011).

Figure 9 shows a possible expansion of reaCog. Two motivation units represent the state “awake” and the state “sleep”, respectively. In the awake state, several sensory and/or motor elements can be activated. These elements may form different contextual groups. To simplify matters, here we focus on two such groups only. One group contains the procedure “grasp” and a memory element representing the visually-given input “position of an object” (relative to the agent), in this case the position of a piece of candy (pos.candy), which is hidden under a cover. We further assume that the agent can also recognize, as a specific kind of object, a conspecific (“partner”), (see Steels & Spranger 2008 and Spranger et al. 2009 for solutions), to whom the agent can attribute properties. These are, in our example, the memory elements “face” and “position”, which stand for the visual appearance and spatial location of the partner to be recognized. Together with the unit “partner” these motivation units form an excitatory network (the dashed connections marked 1A and 1B will be treated later). The procedure “grasp” contains a body-model consisting of an RNN (Schilling 2011) that contains information on the arm used for grasping. This network can be applied to both motor control and recognition of the arm. The former function is symbolized by the output arrow. Concerning the latter function, the body-model is used to minimize errors between the position of the internal model of an arm and the (underspecified) visual input of the arm (e.g., Schilling 2011). If the error could be made small enough, the visual input can be interpreted so as to match the morphology and the specific spatial configuration of the model arm. To symbolize this capability, in figure 9 the procedure “grasp” is also equipped with sensory (visual) input.

The network depicted in figure 9 (disregarding connections 1A and 1B) enables the agent to recognize the position of the candy and to grasp it (“Ego grasp candy”), as indicated by the motivation units marked red in figure 9a. It further allows recognition of the face and the position of the partner. But it does not enable

the agent to “put itself into the partner’s shoes”. In other words, the agent is not able to realize that the partner may have his/her own representation of the world. Thus, the capability of a ToM is lacking.

The motivation unit connecting the agent-related elements “pos.candy” and “grasp” has been called “Ego” in the figures. Although not required for the functioning of this network as shown in figure 9a (disregarding connections 1A and 1B), the application of the unit Ego would allow the introduction of a Word-net representing the word “I”. Thus, with this expansion the concept of “I”, as opposed to other agents (e.g., a partner), can be used by our agent, allowing for internal states like “I grasp candy”, and therefore for self-representation.

Unit Ego is, however, necessary in our framework when two units (here “Ego” and “Partner”) share elements, as will be the case in the following example, where we will enable the agent to represent the partner performing a grasping movement. To this end, we introduce mutual excitatory connections between the unit representing the partner and the procedural element “grasp” (dashed excitatory connection 1A, figure 9). In addition, Unit “Ego” and unit “Partner” have to be connected via mutual inhibition (dashed inhibitory connection 1B, figure 9). This inhibitory connection has the effect that only one of the units—either unit “Ego” or unit “Partner”—can be activated at a given moment in time. With these additional connections 1A and 1B, the network can adopt the internal state “Partner grasp candy”. This situation can be represented in the agent’s memory by activation of the motivation units illustrated in figure 9b, highlighted in red. Note that the introduction of connections 1A and 1B does not alter the ability of the agent to represent the situation “Ego grasp candy” addressed above.

The architecture depicted in figure 9, including connections 1A and 1B, has eventually been termed the application of “shared circuits”, since the procedure “grasp” can be addressed by both unit “Ego” and unit “Partner”, which strongly reminds us of properties characterizing mirror neurons. Therefore, application of such shared circuits has been described as

“mirroring” (Keysers & Gazzola 2007). Units of the grasp-net (including the target pos.candy) represent the movement and its goal, and thus correspond to representations of a motor act, such as has been attributed to mirror neurons (Rizzolatti & Luppino 2001). The grasping movement in both cases (figure 9a, b) is represented as being viewed by the agent (“Ego grasp candy”, figure 9a) or by the partner (“Partner grasp candy”, figure 9b). This means that there is still no ToM possible for the agent. To enable the agent to develop a ToM, we need another expansion.

To explain this, we will present a simple simulation of the Sally–Anne task mentioned above. Both protagonists, Sally and Anne, may have different memory contents concerning the position of the candy. This means that the agent, in this case Anne, needs to be able to represent some aspects of the memory of her partner, too. Therefore, the memory section representing her partner will be equipped with a memory element representing the position of the candy as viewed by her partner Sally, who left the room (figure 10, connection 2). Both memory elements that have possible access to the procedure “grasp” have to be connected by mutual inhibition, so that only one of these elements can address the procedure at a given time in order to allow for sensible representation of the situation. Now imagine that the subject Anne is either equipped with a network as depicted in figure 9, or that depicted in figure 10. Application of a system as shown in figure 9 means that the agent (Anne) has only one representation of the candy’s position, namely the one seen last. Therefore only this, correct, position can be activated and it is imagined that the partner grasps the correct position—this kind of prediction is observed in children younger than about four years. Anne cannot take into account the likely assumption her partner will make about the location of the candy. In contrast, in a system as presented in figure 10, there is a difference in thinking of oneself grasping the candy or the partner grasping it. When the agent, Anne, imagines herself grasping the candy, she would grasp its position as under the correct cover (figure 10a). If asked

to simulate the internal state of her partner, as is required in the case of the Sally–Anne test (figure 10b), the position connected to her partner Sally is used and the agent will rightfully deduct that her partner’s grasp would be directed towards this position—which is wrong, but this fact is not known by her partner. Therefore, the network shown in figure 10 allows for ToM, in contrast to the network shown in figure 9. The critical difference between both networks is that the network shown in figure 10 contains a separate representation of (a part of) the partner’s memory. This means that a comparat-

11 Discussion

Consciousness and the relation of the outside world to mental representation are central to philosophy of mind, and have led to many diverse views (Vision 2011). While many of those views appear plausible in themselves, especially from a non-philosopher’s perspective, there appears to be much disagreement among philo-

sophers. Many of the positions are based on high-level views approaching consciousness in a top-down fashion. In contrast, our approach starts from a low-level control system for a behaving agent. The goal is the bottom-up development of higher-level faculties. In this way, the neural architecture implements a minimal cognitive system that can be used as a hypothesis for cognitive mechanisms and higher-level functioning, which are testable in a real-world system, for example, on a robot. This allows deriving testable and quantitative hypotheses for higher-level phenomena. In this way, a bottom-up approach can nicely complement philosophical discussions focusing mainly on higher-level aspects. In addition, such a minimal cognitive system can provide functional descriptions of higher-level properties. We briefly introduced the reaCog system in this article, following this bottom-up approach. The central concern is the emergent properties that can be identified when analyzing this system. In particular, high-level properties, such as emotions, attention, intention, volition, or consciousness have been considered here and related to the system.

From our point of view, such a bottom-up approach leads to a system that can be used to test quantitative hypotheses. Even though the system was not intended to model, for example, consciousness, the system can be thoroughly analyzed and emergent properties can be related to mental phenomena. This is particularly interesting, as high-level descriptions can leave a lot of room for interpretation. In contrast, connecting mental phenomena to mechanisms of a well-defined system allows for detailed studies and clear-cut definitions on a functional level. In this way, a system can be examined with respect to many even diverging views and may allow resolving ambiguities. Knowledge gained from analyzing the system can in this way inform philosophical theories and refine existing definitions by defining sufficient aspects as well as missing criteria.

One might ask if higher-level phenomena as considered here are not simply too far removed for such a simple system. One basic problem is represented by the frequently-formulated assumption that all these phenomena have

to be tied to the notion of an internal perspective and that phenomenality has a function in and of itself. In contrast, we claim that focusing on the functional aspect is a sensible approach. It is possible because we believe that the phenomenal aspect is always coupled to specific, yet unknown, properties of the neuronal system that, at the same time, have functional effects and show subjective experience. In other words, adopting a monist view, we assume that we can circumvent the “hard” problem, i.e., the question concerning the subjective aspect of mental phenomena, without losing information concerning the function of the underlying procedures. Of course, we are not in a position to claim which of these structures, if any, are accompanied by phenomenality. If, however, the function of, for example, the artificial system indeed corresponds well enough to those of the neuronal structures that are accompanied by phenomenality, the artificial system may have this property, too.

The control network reaCog consists of local procedural modules. We have presented two subnetworks: Walknet, which aims at the control of walking, and Navinet, which deals with navigation. Both consist of a heterarchical structure of motivation units that form a recurrent neural network. This, via competition and cooperation between those units, allows for various attractor states that enforce action selection. Selection of one or a group of procedures protects a current behavioral context against non-relevant sensory input. An internal model of the body is part of the control network coordinating joint movements in walking. As this model is quite flexible and predictive, it can be used for planning ahead through internal simulation. Following the definition of [McFarland & Bösner \(1993\)](#), the network, since it is based on reactive procedures and is capable of planning ahead, can be termed a cognitive system, giving rise to its name: reaCog. In combination with the attention controller, the whole framework can come up with new behavioral solutions when encountering problems, i.e., behaviors that are not automatically activated by the current context. Internal simulation allows us to test these behaviors and to come up with pre-

dicted consequences, which can be used to guide the selection process for the real system. The attention controller cannot function independently. It is tightly connected to the reactive structures. The procedural memory of the reactive system is further accompanied with perceptual memory and Word-nets, a specific form of mixed procedural and perceptual memory. The latter memory elements allow the introduction of symbolic information. Symbol-grounding is realized by specific connections between the motivation unit of a Word-net and its partner motivation unit, representing the corresponding concept in the procedural (or the perceptual) memory.

Key characteristics of reaCog are modularity, heterarchy, redundancy, cross-modal influences (e.g., path integration and landmark navigation in Navinet), bottom-up and top-down attention control, i.e., the selection of relevant sensory inputs, as well as recruitment of internal models for planning. The complete control system constitutes a holistic system as the central selection control process—including the internal body-model—is implemented as an RNN. Overall, reaCog follows Anderson’s massive redeployment hypothesis (Anderson 2010), since large parts of the reactive control network structure are reused in higher-level tasks (as discussed in detail in section 4 for planning ahead and in section 10.3 for Theory of Mind).

ReaCog nicely demonstrates how complex behavior can emerge from the interaction of simple control networks and coordination on a local level, as well as through the loop through the environment. Its feasibility is shown through the implementation of the system at first in dynamic simulation (for Navinet on a two DoF, wheeled robot platform; for Walknet using a hexapod, twenty-two DoF hexapod robot). Second, those control networks are currently applied to a real robot, called Hector (Schneider et al. 2011).

Emergent properties are properties that are to be addressed using levels of description other than those used to describe the properties of the elements. In the reactive part of the system (Walknet, Navinet) we have already found

some emergent properties (development of different “gaits”, climbing over large gaps, finding shortcuts in navigation characterized as cognitive-map-like behavior) as well as forms of bottom-up and top-down attention. With respect to the notion of access consciousness, several contributing properties are present in reaCog. Most notably, planning ahead through internal simulation is central to reaCog. New behavioral plans are tested in the internal simulation, thus exploiting the existing internal model and its predictive capabilities. Only afterwards are successful behaviors applied on the real agent. In this way, the agent can deal with novel contexts and is not restricted to the hard-wired structure of the reactive system.

Furthermore, the system shows global availability, which means that elements of the procedural memory can be addressed even if they do not belong to the current context. A third property contributing to elements forming access consciousness concerns the ability of the system to communicate with an external supervisor by following (i.e., understanding) verbal commands and by reporting on its internal states. Therefore, except for the ability of linguistic reasoning, which is clearly missing, the issues characterizing access consciousness as listed by Cleeremans (2005) are fulfilled. But there are also disadvantages: (i) First, reactive automatic control is faster. As cognitive control involves internal simulation (and probably multiple simulations) the whole process takes more time. In addition, there is an overhead of higher-level control going on in contrast to reactive control. (ii) While access consciousness enables the system to deal with novel situations and to come up with new behaviors, the same processes might interfere when they are active during processing of the reactive control level. This might lead to worse performance when both levels are active at the same time. Both mentioned drawbacks have been confirmed in psychological experiments. We have not dealt with the subjective aspect of consciousness. But leaving this aside, we have shown how reaCog shows important constituent properties of access consciousness and how it may provide, in this way, a scaffold for a more complex system that

can manifest additional basic aspects of consciousness.

The property of having an internal body-model and the property of being able to internally simulate behavior have been explicitly implemented and can therefore not be considered emergent properties in our approach. However, when referring to a hypothetical evolutionary process that may have led to the development of these properties, the appearance of the body-model and of cognitive expansion might well be characterized as representing an emergent property.

We based our analysis and discussion on the perspective of Cleeremans, and used his concepts. One counter argument addressing the notion of access consciousness is that this notion is too unspecific as it does not help to distinguish between systems, and may cover “too many” systems. For instance, one may ask, following a minimalist approach, whether this notion of access consciousness might even include programs like chess-playing software. One might also ask whether there is a fundamental difference between such a system and a system like *reaCog*.

While both systems are able to search for the solution to a problem using internal simulation, there are indeed crucial differences. A typical chess program would be not embodied, but, obviously, today this difference can be easily overcome and the system could be realized in a robot equipped with a vision system and a hand that could move the chess figures.

However, more importantly, the basic difference between such a chess player and *reaCog* would be their flexibility in using internal models. A chess-playing robot always operates within the same context, which is stored in a separate memory-domain, for example in a list of symbolic rules. In contrast, *reaCog* basically operates with a reactive system, but can also switch to the state of internal simulation when a problem occurs. It then searches for a solution by testing memory elements not belonging to the actual context. In other words, *reaCog* is able to exchange information between different contextual domains. Such a switch is not available to a chess-playing program at all. Such a

program cannot distinguish between different contexts. In other words, there is no global accessibility in the sense described for systems showing access consciousness. As a consequence, the discussion of drawbacks connected with access consciousness as mentioned in the above paragraph on emergent properties, that is, issues (i), and (ii), is not applicable to such a chess-playing system, and nor are the dynamical effects observed in the experiments of [Beilock et al. \(2002\)](#) 1 (section 10.2).

The same holds for the phenomena of a psychological refractory period, attentional blink, and the masking experiments discussed earlier in section 7. None of these phenomena can be addressed by a classical chess-player system, first, because due to the different architectures, no search of a domain belonging to a different context is possible. A chess player does not meet the requirements of access consciousness as listed by [Cleeremans 2005](#) and represented by *reaCog*. Second, no specific dynamics can be found in such a chess-player system that could be made responsible for the dynamical effects mentioned above and which may provide the substrate for the occurrence of phenomenal experience. Therefore, both systems are qualitatively different. If at all, the chess player may correspond to a subsection of the symbolic domain of access consciousness, which has not yet been explicitly addressed in this article.

In an earlier paper ([Cruse & Schilling 2013](#)), taking a conservative position, we argued that properties of metacognition could not be found in the earlier version of *reaCog*. We have now provided some new arguments that permit a different position concerning this matter. Using this architecture, the agent is able to monitor internal states and use this information to control its behavior. Internal states may also be able to represent the agent itself. A first expansion allows representation of the activations of a partner by using the same procedure as is used for controlling the agent’s own behavior (application of “shared” circuits, “mirroring”). Furthermore, using an expansion proposed by [Cruse & Schilling \(2011\)](#), the agent is also able to exploit and represent knowledge about the internal states of others, specifically by applying ToM.

Cruse & Schilling (2011) have further shown how this network can be expanded to represent the discrimination between subject and object (e.g., Ego push Partner) and to attribute subjective experience (e.g., pain) to the partner using a shared body-model. A further expansion that allows for mutualism—two agents cooperate to reach a common goal (“shared intention”, Tomasello 2009)—requires two body-models, corresponding to what Tomasello calls a we-model.

In the remainder of this section we briefly mention some aspects not addressed by reaCog. First, not all combinations of the elements explained for our network have been tested within the complete system. For example, Walknet and Navinet have been tested in separate software and hardware simulations. Second, we concentrated on solving motor problems alone, and did not deal with how this system could solve problems in the symbolic domain at all. From an embodied point of view, this restriction is not as problematic as it might initially seem, as the solution process for many problems can be traced back to abilities that are based on solving motor tasks (Glenberg & Gallese 2011); for example this holds true even for abstract domains such as mathematical problem-solving (Lakoff & Nunez 2000).

Finally, an important aspect not addressed here in any detail concerns how learning of the memory elements, including the weight of the motivation unit network, is possible. Examples of learning position and quality of new food sources in Navinet are given by Hoinville et al. (2012), examples of learning perceptual networks, including the heterarchical arrangement of concepts, are given by Cruse & Schilling (2010a), but introduction of the ability to learn such properties within the complete system has not yet been introduced.

12 Conclusion

We describe a way to construct an artificial agent whose architecture is characterized by a number of local, reactive procedures controlled by an RNN, termed motivation unit network. This network is able to adopt various attractor

states, or internal states, which are able to protect the complete system from sensory input not belonging to the current internal state. No strict hierarchy can be observed in this network. Instead, internal states may be represented by partly overlapping state vectors.

Where required, further procedures have been introduced that can be interpreted as forming explicit representations of parts of the environment. Specifically, an internal model of the agent’s own body is introduced that can, as a “manipulable” body-model, be used for planning new behaviors via internal simulation. Internal manipulation is possible because the body-model, like a marionette puppet, able to adopt all configurations the real body can assume. This expansion allows the agent to switch between reactive control and cognitive control (in the sense of McFarland & Bösner 1993).

When aiming to study higher mental properties, at least in human beings, we have to deal with the phenomenal aspect of these properties. A number of experimental results suggest that, i) some, but not all neuronal activities are, under specific—and unknown in any detail—conditions equipped with a phenomenal aspect, i.e., show subjective experience, but that ii) there is no specific function of this phenomenal aspect apart from the functions that can be ascribed to the physical properties of the system. Note that this does not mean that the phenomenal aspect has no function. Rather, a network adopts the function only when, at the same time, the phenomenal aspect is given. This view allows us to focus the analysis on the functional aspect of the procedure (see section 7). However, due to our lack of knowledge, as an external observer we cannot decide whether a given internal state is a mental state or not (if mental states are understood as internal states that are equipped with a phenomenal aspect).

The complete network represents a collection of hypotheses that can be tested by comparing their properties with experimental data and by trying to match them with theoretical concepts. Examples studied in this article concern behaviors that, for an external observer, may be conceptualized as various gait patterns, or navigation using an internal map, on the

“lower” level. On a higher level, we deal with inventing new behaviors and planning ahead, as well as phenomena attributed to mental states like emotions, attention, intention, and volition. Last but not least we compare the properties of our approach with different aspects of consciousness, such as access consciousness (including global accessibility) and metacognition. We claim that, at least in their basic form, these phenomena can be attributed to internal states emerging from the cooperation of decentralized elements of our network.

References

- Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences*, 33 (4), 254-313. [10.1017/S0140525X10000853](https://doi.org/10.1017/S0140525X10000853)
- Baars, B. J. & Franklin, S. (2007). An architectural model of conscious and unconscious brain functions: Global workspace theory and IDA. *Neural Networks*, 20 (9), 955 - 961. [10.1016/j.neunet.2007.09.013](https://doi.org/10.1016/j.neunet.2007.09.013)
- Beilock, S. L., Carr, T. H., MacMahon, C. & Starkes, J. L. (2002). When paying attention becomes counter-productive: Impact of divided versus skill-focussed attention on novice and experienced performance of sensorimotor skills. *Journal of Experimental Psychology: Applied*, 8 (1), 6-16. [10.1037/1076-898X.8.1.6](https://doi.org/10.1037/1076-898X.8.1.6)
- Bloch, A. M. (1885). Expérience sur la vision. *Comptes Rendus de Séances de la Société de Biologie*, 37, 493-495.
- Block, N. (1995). On a confusion about a function of consciousness. *The Behavioral and Brain Sciences*, 18 (2), 227-287. [10.1017/S0140525X00038188](https://doi.org/10.1017/S0140525X00038188)
- (2001). Paradox and cross purposes in recent work on consciousness. *Cognition*, 79 (1-2), 197-219. [10.1016/S0010-0277\(00\)00129-3](https://doi.org/10.1016/S0010-0277(00)00129-3)
- Bläsing, B. (2006). Crossing large gaps: A simulation study of stick insect behavior. *Adaptive Behavior*, 14 (3), 265-285. [10.1177/105971230601400307](https://doi.org/10.1177/105971230601400307)
- Bratman, M. E. (1987). *Intention, plans and practical reason*. Cambridge, MA: Harvard University Press.
- Carruthers, P. (1996). Simulation and self-knowledge: A defence of the theory-theory. In Carruthers, P. and Smith, P.K. (Eds.) *Theories of theories of mind*. Cambridge, UK: Cambridge University Press.
- Chalmers, D. (1997). *The conscious mind : In search of a fundamental theory*. New York, NY: Oxford University Press.
- Chittka, L. & Niven, J. (2009). Are bigger brains better? *Current Biology* (19), R995-R1008. [10.1016/j.cub.2009.08.023](https://doi.org/10.1016/j.cub.2009.08.023)
- Cleeremans, A. (2005). Computational correlates of consciousness. *Progress in Brain Research*, 150, 81-98. [10.1016/S0079-6123\(05\)50007-4](https://doi.org/10.1016/S0079-6123(05)50007-4)
- Cleeremans, A., Timmermans, B. & Pasquali, A. (2007). Consciousness and metarepresentation: A computational sketch. *Neural Networks*, 20 (9), 1032-1039. [10.1016/j.neunet.2007.09.011](https://doi.org/10.1016/j.neunet.2007.09.011)
- Cohen, M. A. & Dennett, D. C. (2011). Consciousness cannot be separated from function. *Trends in Cognitive Sciences*, 15 (8), 358-64. [10.1016/j.tics.2011.06.008](https://doi.org/10.1016/j.tics.2011.06.008)
- Cruse, H. (2003). The evolution of cognition: A hypothesis. *Cognitive Science*, 27 (1), 135-155. [10.1207/s15516709cog2701_5](https://doi.org/10.1207/s15516709cog2701_5)
- Cruse, H. & Schilling, M. (2010a). Learning and retrieval of hierarchically organized information in a simple, one-layered RNN. *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2010 at WCCI 2010 IEEE World Congress on Computational Intelligence)*, Barcelona, Spain (pp. 521-528). [10.1109/IJCNN.2010.5596804](https://doi.org/10.1109/IJCNN.2010.5596804)
- (2010b). Getting cognitive. In Bläsing, P., Puttke, M. and Schack, T. (Eds.) *The Neurocognition of Dance* (pp. 53-74). Psychology Press.
- (2011). From egocentric systems to systems allowing for theory of mind and mutualism. In R. Doursat (Ed.) *Proceedings of the ECAL 2011, Paris* (pp. 184-191). MIT Press.
- (2013). How and to what end may consciousness contribute to action? Attributing properties of consciousness to an embodied, minimally cognitive artificial neural network. *Frontiers in Psychology*, 4 (324). [10.3389/fpsyg.2013.00324](https://doi.org/10.3389/fpsyg.2013.00324)
- (2014). Action selection within short time windows. *Biomimetic and Biohybrid Systems*, 8608, 47-58. [10.1007/978-3-319-09435-9_5](https://doi.org/10.1007/978-3-319-09435-9_5)
- Cruse, H. & Wehner, R. (2011). No need for a cognitive map: Decentralized memory for insect navigation. *PLoS Computational Biology*, 7 (3), e1002009. [10.1371/journal.pcbi.1002009](https://doi.org/10.1371/journal.pcbi.1002009)
- Dehaene, S. & Changeux, J. P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, 70 (2), 200-227. [10.1016/j.neuron.2011.03.018](https://doi.org/10.1016/j.neuron.2011.03.018)
- Dehaene, S. & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, 79 (1-2), 1-37. [10.1016/S0010-0277\(00\)00123-2](https://doi.org/10.1016/S0010-0277(00)00123-2)

- Dennett, D. C. (1991). *Consciousness explained*. Boston, MA: Little, Brown & Co.
- Desimone, R. & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18, 193-222. [10.1146/annurev.ne.18.030195.001205](https://doi.org/10.1146/annurev.ne.18.030195.001205)
- Doya, K. (2000). Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current Opinion in Neurobiology*, 10 (6), 732-739. [10.1016/S0959-4388\(00\)00153-7](https://doi.org/10.1016/S0959-4388(00)00153-7)
- (2002). Metalearning and neuromodulation. *Neural Networks*, 15 (4-6), 495-506. [10.3410/f.1001684.173108](https://doi.org/10.3410/f.1001684.173108)
- Dreisbach, G. & Goschke, T. (2004). How positive affect modulates cognitive control: reduced perseveration at the cost of increased distractibility. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30 (2), 343-53. [10.1111/j.1460-9568.2007.05949.x](https://doi.org/10.1111/j.1460-9568.2007.05949.x)
- Dürr V., Schmitz, J. & Cruse, H. (2004). Behaviour-based modelling of hexapod locomotion: Linking biology and technical application. *Arthropod Structure & Development*, 33 (3), 237-250. [10.1016/j.asd.2004.05.004](https://doi.org/10.1016/j.asd.2004.05.004)
- Ekman, P. (1999). Basic emotions. In Dalglish, T. and Power, M. J. (Eds.) *Basic emotions* (pp. 45-60). New York, NY: John Wiley & Sons Ltd..
- Fehrer, E. & Raab, D. (1962). Reaction time to stimuli masked by metacontrast. *Journal of Experimental Psychology*, 62 (2), 143-147. [10.1037/h0040795](https://doi.org/10.1037/h0040795)
- Fossat, P., Bacqué-Cazenave, J., De Deurwaerdère, P., Delbecq, J.-P. & Cattaert, D. (2014). Comparative behavior. Anxiety-like behavior in crayfish is controlled by serotonin. *Science*, 344 (6189), 1293-1297. [10.1126/science.1248811](https://doi.org/10.1126/science.1248811)
- Gibson, J. J. (1979). *The ecological approach to visual perception*. New Jersey: Lawrence Erlbaum Associates.
- Glenberg, A. M. (1997). What memory is for. *Behavioral and Brain Sciences*, 20 (1), 1-55.
- Glenberg, A. M. & Gallese, V. (2011). Action-based language: A theory of language acquisition, comprehension, and production. *Cortex*, 48 (7), 905-922. [10.1016/j.cortex.2011.04.010](https://doi.org/10.1016/j.cortex.2011.04.010)
- Goldman, A. (2005). Imitation, mind reading, and simulation. In Hurley, S. and Chater, N. (Eds.) *Perspectives on imitation II* (pp. 80-81). Cambridge, MA: MIT Press.
- Goschke, T. (2013). Volition in action: intentions, control dilemmas, and the dynamic regulation of cognitive control. In Prinz, W., Beisert, M. and Herwig, A. (Eds.) *Action science: Foundations of an emerging discipline* (pp. 409-434). Cambridge, MA: MIT Press.
- Hesslow, G. (2002). Conscious thought as simulation of behaviour and perception. *Trends in Cognitive Sciences*, 6 (6), 242-247. [10.1016/s1364-6613\(02\)01913-7](https://doi.org/10.1016/s1364-6613(02)01913-7)
- Hoinville, T., Wehner, R. & Cruse, H. (2012). Learning and retrieval of memory elements in a navigation task. In Prescott, T.J., Lepora, N.F., Mura, A. and Verschure, P.F.M.J. (Eds.) *Biomimetic and Biohybrid Systems* (pp. 120-131). [10.1007/978-3-642-31525-1_11](https://doi.org/10.1007/978-3-642-31525-1_11)
- Holland, O. & Goodman, R. (2003). Robots with internal models: A route to machine consciousness? *Journal of Consciousness Studies, Special Issue on Machine Consciousness*, 10 (4-5), 77-109.
- Keysers, C. & Gazzola, V. (2007). Integrating simulation and theory of mind: From self to social cognition. *Trends in Cognitive Sciences*, 11 (5), 194-196. [10.1016/j.tics.2007.02.002](https://doi.org/10.1016/j.tics.2007.02.002)
- Koch, C. & Tsuchiya, N. (2007). Attention and consciousness: Two distinct brain processes. *Trends in Cognitive Sciences*, 11 (1), 16-22. [10.1016/j.tics.2006.10.012](https://doi.org/10.1016/j.tics.2006.10.012)
- Kouider, S., de Gardelle, V., Sackur, J. & Dupoux, E. (2010). How rich is consciousness? The partial awareness hypothesis. *Trends in Cognitive Sciences*, 14 (7), 301-307. [10.1016/j.tics.2010.04.006](https://doi.org/10.1016/j.tics.2010.04.006)
- Kugler, P. N., Shaw, R. E., Vicente, K. J. & Kinsella-Shaw, J. (1990). Inquiry into intentional systems I: Issues in ecological physics. *Psychol. Res.*, 52 (2), 98-121. [10.1007/BF00877518](https://doi.org/10.1007/BF00877518)
- Lakoff, G. & Nunez, R. (2000). *Where mathematics comes from: How the embodied mind brings mathematics into being*. New York, NY: Basic Books.
- Laughlin, R. B. & Pines, D. (2000). The theory of everything. *Proceedings of the National Academy of Sciences of the United States of America*, 97 (1), 28-31. [10.1073/pnas.97.1.28](https://doi.org/10.1073/pnas.97.1.28)
- Lau, H. & Rosenthal, D. M. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*, 15 (8), 365-373. [10.1016/j.tics.2011.05.009](https://doi.org/10.1016/j.tics.2011.05.009)
- Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, 64, 354-361.
- Libet, B., Alberts, W. W., Wright, E. W., Delattre, L. D., Levin, G. & Feinstein, B. (1964). Production of threshold levels of conscious sensation by electrical stimulation of human somatosensory cortex. *Journal of Neurophysiology*, 27 (4), 546-578. [10.1007/978-1-4612-0355-1_1](https://doi.org/10.1007/978-1-4612-0355-1_1)
- Loeb, G. E. (2001). Learning from the spinal cord. *Journal of Physiology*, 533, 111-117. [10.1111/j.1469-7793.2001.0111b.x](https://doi.org/10.1111/j.1469-7793.2001.0111b.x)

- Maes, P. (1991). A bottom-up mechanism for behavior selection in an artificial creature. In Meyer, J.-A. and Wilson, S.W. (Eds.) (pp. 238-246). Cambridge, MA: MIT Press.
- McFarland, D. & Bösner, T. (1993). *Intelligent behavior in animals and robots*. Cambridge, MA: MIT Press.
- Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 4, 261-292.
- Menzel, R., Brembs, B. & Giurfa, M. (2007). Cognition in invertebrates. In Kaas, J.H. (Ed.) *Evolution of nervous systems in invertebrates* (pp. 403-442). Oxford, UK: Oxford University Press.
- Metzinger, T. (2006). Different conceptions of embodiment. *Psyche*, 12 (4)
- (2009). *The ego tunnel - The science of the mind and the myth of the self*. New York, NY: Basic Books.
- (2013). Two principles for robot ethics. In Hilgendorf, E. and Günther, J.-P. (Eds.) *Robotik und Gesetzgebung* (pp. 263-302). Baden-Baden: Nomos.
- (2014). First-order embodiment, second-order embodiment, third-order embodiment: From spatiotemporal self-location to minimal phenomenal selfhood. In Shapiro, L.A. (Ed.) *The Routledge Handbook of Embodied Cognition* (pp. 272-286). London, UK: Routledge.
- Mussa-Ivaldi, F. A., Morasso, P. & Zaccaria, R. (1988). Kinematic networks distributed model for representing and regularizing motor redundancy. *Biological Cybernetics*, 60 (1), 1-16. [10.1007/BF00205967](https://doi.org/10.1007/BF00205967)
- Narayanan, S. (1997). Talking the talk is like walking the walk: A computational model of verbal aspect. *COGSCI-97* (pp. 548-553). Stanford, CA. [10.1.1.35.1211](https://doi.org/10.1.1.35.1211)
- Neisser, U. (1967). *Cognitive psychology*. New York, NY: Appleton-Century-Crofts.
- Neumann, O. & Klotz, W. (1994). Motor responses to non-reportable, masked stimuli: Where is the limit of direct parameter specification? In Umiltà, C. and Moscovitch, M. (Eds.) *Attention and performance XV* (pp. 123-150). Cambridge, MA: MIT Press.
- Norman, D. A. & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In Davidson, R.J., Schwartz, G.E. and Shapiro, D. (Eds.) *Consciousness and self-regulation: Advances in research and theory* (pp. 1-18). New York, NY: Plenum.
- O'Connor, C. M., Cree, G.S. & McRae, K. (2009). Conceptual hierarchies in a flat attractor network: Dynamics of learning and computations. *Cognitive Science*, 33 (4), 665-708. [10.1111/j.1551-6709.2009.01024.x](https://doi.org/10.1111/j.1551-6709.2009.01024.x)
- Pacherie, E. (2006). Toward a dynamic theory of intentions. In Pockett, S., Banks W.P. and Gallagher, S. (Eds.) *Does consciousness cause behavior? An investigation of the nature of volition* (pp. 145-167). Cambridge, MA: MIT Press.
- Parisi, D. & Petrosino, G. (2010). Robots that have emotions. *Adaptive Behavior*, 18 (6), 453-469. [10.1177/1059712310388528](https://doi.org/10.1177/1059712310388528)
- Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In Plutchik, R. and Kellerman, H. (Eds.) *Emotion: Theory, research, and experience: Vol. 1. Theories of emotion* (pp. 3-23). New York, NY: Academic.
- Premack, D. G. & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral Brain Sciences*, 1 (4), 515-526. [10.1017/S0140525X00076512](https://doi.org/10.1017/S0140525X00076512)
- Pérez, C. H., Escibano, G. S. & Sanz, R. (2012). The morphological approach to emotion modelling in robotics. *Adaptive Behavior*, 20 (5), 388-404. [10.1177/1059712312451604](https://doi.org/10.1177/1059712312451604)
- Rizzolatti, G. & Luppino, G. (2001). The cortical motor system. *Neuron*, 31 (6), 889 - 901. [10.1016/S0896-6273\(01\)00423-8](https://doi.org/10.1016/S0896-6273(01)00423-8)
- Rosenthal, D. M. (2002). How many kinds of consciousness? *Consciousness and Cognition*, 11 (4), 653-665. [10.1016/S1053-8100\(02\)00017-X](https://doi.org/10.1016/S1053-8100(02)00017-X)
- Russell, S. & Norvig, P. (2003). *Artificial intelligence: A modern approach*. Englewood Cliffs, NJ: Prentice-Hall.
- Schier, E. (2009). Identifying phenomenal consciousness. *Consciousness and Cognition*, 18 (1), 216-222. [10.1016/j.concog.2008.04.001](https://doi.org/10.1016/j.concog.2008.04.001)
- Schilling, M. (2011). Universally manipulable body models - Dual quaternion representations in layered and dynamic (MMCs). *Autonomous Robots*, 30 (4), 399-425. [10.1007/s10514-011-9226-3](https://doi.org/10.1007/s10514-011-9226-3)
- Schilling, M. & Cruse, H. (2007). Hierarchical MMC networks as a manipulable body model. *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2007)*, Orlando, FL (pp. 2141-2146). Orlando, FL. [10.1109/IJCNN.2007.4371289](https://doi.org/10.1109/IJCNN.2007.4371289)
- (2008). The evolution of cognition - From first order to second order embodiment. In Wachsmuth, I. and Knoblich, G. (Eds.) *Modeling Communication with Robots and Virtual Humans* (pp. 77-108). Springer, GER: Springer.
- (2012). What's next: Recruitment of a grounded predictive body model for planning a robot's actions. *Frontiers in Cognition*, 3 (383), [10.3389/fpsyg.2012.00383](https://doi.org/10.3389/fpsyg.2012.00383)
- Schilling, M. & Cruse, H. (submitted). reaCog, a minimal cognitive controller based on recruitment of reactive systems.

- Schilling, M., Schneider, A., Cruse, H. & Schmitz, J. (2008). Local control mechanisms in six-legged walking. *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2008* (pp. 2655-2660). [10.1109/iros.2008.4650591](https://doi.org/10.1109/iros.2008.4650591)
- Schilling, M., Paskarbit, J., Schmitz, J., Schneider, A. & Cruse, H. (2012). Grounding an internal body model of a hexapod walker - Control of curve walking in a biological inspired robot. *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2012* (pp. 2762-2768). [10.1109/iros.2012.6385709](https://doi.org/10.1109/iros.2012.6385709)
- Schilling, M., Hoinville, T., Schmitz, J. & Cruse, H. (2013a). Walknet, a bio-inspired controller for hexapod walking. *Biological Cybernetics*, 107 (4), 397-419. [10.1007/s00422-013-0563-5](https://doi.org/10.1007/s00422-013-0563-5)
- Schilling, M., Paskarbit, J., Hoinville, T., Hüffmeier, A., Schneider, A., Schmitz, J. & Cruse, H. (2013b). A hexapod walker using a heterarchical architecture for action selection. *Frontiers in Computational Neuroscience*, 7. [10.3389/fncom.2013.00126](https://doi.org/10.3389/fncom.2013.00126)
- Schmitz, J., Schneider, A., Schilling, M. & Cruse, H. (2008). No need for a body model: Positive velocity feedback for the control of an 18DOF robot walker. *Applied Bionics and Biomechanics*, 5 (3), 135-147. [10.1080/11762320802221074](https://doi.org/10.1080/11762320802221074)
- Schneider, W. (2013). Selective visual processing across competition episodes: a theory of task-driven visual attention and working memory. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 368 (1628), 20130060. [10.1098/rstb.2013.0060](https://doi.org/10.1098/rstb.2013.0060)
- Schneider, A., Paskarbit, J., Schäffersmann, M. & Schmitz, J. (2011). Biomechatronics for of embodied intelligence an insectoid robot. *Proc. ICRA 2* (pp. 1-11). [10.1007/978-3-642-25489-5_1](https://doi.org/10.1007/978-3-642-25489-5_1)
- Seth, A. (2007). Models of consciousness. *Scholarpedia* 2, 1328, 2 (1), 1328. [10.4249/scholarpedia.1328](https://doi.org/10.4249/scholarpedia.1328)
- Singer, W. & Gray, C. M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neuroscience*, 18, 555-586. [10.1146/annurev.ne.18.030195.003011](https://doi.org/10.1146/annurev.ne.18.030195.003011)
- Spranger, M., Höfer, S. & Hild, M. (2009). Biologically inspired posture recognition and posture change detection for humanoid robots. *Proceedings of ROBIO'09: IEEE International Conference on Robotics and Biomimetics*. (pp. 562-567). [10.1109/ROBIO.2009.5420708](https://doi.org/10.1109/ROBIO.2009.5420708)
- Steels, L. (1995). Intelligence - Dynamics and Representations. In Steels, L. (Ed.) (pp. 72-89). New York, NY: Springer. [10.1007/978-3-642-79629-6_4](https://doi.org/10.1007/978-3-642-79629-6_4)
- (2003). Intelligence with representation. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, 361 (1811), 2381-2395. [10.1098/rsta.2003.1257](https://doi.org/10.1098/rsta.2003.1257)
- (2007). The symbol grounding problem is solved, so what's next? In De Vega, M., Glennberg, G. and Graesser, G. (Eds.) *Symbols, embodiment and meaning*. New Haven, CT: Academic Press.
- Steels, L. & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, 28 (04), 469-489. [10.1017/S0140525X05000087](https://doi.org/10.1017/S0140525X05000087)
- Steels, L. & Spranger, M. (2008). The robot in the mirror. *Connection Science*, 20 (4), 337-358. [10.1080/09540090802413186](https://doi.org/10.1080/09540090802413186)
- Tomasello, M. (2009). *Why we cooperate*. Cambridge, MA: MIT Press.
- Valdez, P. & Mehrabian, A. (1994). Effects of color on emotions. *Journal of Experimental Psychology: General*, 123 (4), 394-409. [10.1037/a0031821](https://doi.org/10.1037/a0031821)
- Vision, G. (2011). *Re-emergence. Locating conscious properties in a material world*. Cambridge, MA: MIT Press.
- von Kleist, H. (1810). Über das Marionettentheater. In Sembdner, H. (Ed.) *Heinrich von Kleist, Sämtliche Werke und Briefe, Bd. 2* (p. 345). München, GER: Deutscher Taschenbuch Verlag (originally appeared in *Berliner Abendblätter*, 1. Jg., 1810).
- Wundt, W. (1863). *Vorlesung über die Menschen- und Tierseele*. Leipzig, GER: Voss Verlag.
- Yang, Z., Bertolucci, F., Wolf R. & Heisenberg M. (2014). Flies cope with uncontrollable stress by learned helplessness. *Current Biology*, 23 (9), 799-803. [10.1016/j.cub.2013.03.054](https://doi.org/10.1016/j.cub.2013.03.054)
- Zylberberg, A., Dehaene, S., Roelfsema, P. R. & Sigman, M. (2011). The human turing machine: A neural framework for mental programs. *Trends in Cognitive Sciences*, 15 (7), 293-300. [10.1016/j.tics.2011.05.007](https://doi.org/10.1016/j.tics.2011.05.007)

The “Bottom-Up” Approach to Mental Life

A Commentary on Holk Cruse & Malte Schilling

Aaron Gutknecht

With their “bottom-up” approach, Holk Cruse and Malte Schilling present a highly intriguing perspective on those mental phenomena that have fascinated humankind since ancient times. Among them are those aspects of our inner lives that are at the same time most salient and yet most elusive: we are conscious beings with complex emotions, thinking and acting in pursuit of various goals. Starting with, from a biological point of view, very basic abilities, such as the ability to move and navigate in an unpredictable environment, Cruse & Schilling have developed, step-by-step, a robotic system with the ability to plan future actions and, to a limited extent, to verbally report on its own internal states. The authors then offer a compelling argument that their system exhibits aspects of various higher-level mental phenomena such as emotion, attention, intention, volition, and even consciousness.

The scientific investigation of the mind is faced with intricate problems at a very fundamental, methodological level. Not only is there a good deal of conceptual vagueness and uncertainty as to what the explananda precisely are, but it is also unclear what the best strategy might be for addressing the phenomena of interest. Cruse & Schilling’s bio-robotic “bottom-up” approach is designed to provide answers to such questions. In this commentary, I begin, in the first section, by presenting the main ideas behind this approach as I understand them. In the second section, I turn to an examination of its scope and limits. Specifically, I will suggest a set of constraints on good explanations based on the bottom-up approach. What criteria do such explanations have to meet in order to be of real scientific value? I maintain that there are essentially three such criteria: biological plausibility, adequate matching criteria, and transparency. Finally, in the third section, I offer directions for future research, as Cruse & Schilling’s bottom-up approach is well suited to provide new insights in the domain of social cognition and to explain its relation to phenomena such as language, emotion, and self.

Keywords

Bio-robotics | Bottom-up approach | Emergence | Evolution | Explanation | Mechanisms | Robotics | Social cognition

1 Biorobotics and the bottom-up approach to mental life

From my perspective, there are two basic ideas underlying the overall research strategy entertained by Cruse and Schilling. The first is that in order to understand a system and its properties, it has to be *reinvented* or *reconstructed* by the researcher. The second is that mental phe-

nomena may arise as *emergent* properties via the interaction of low-level components of a system. I’d like to first provide an outline of these basic ideas and the underlying strategy as I understand them. In the next section, I will critically evaluate what types of questions the ap-

Commentator

[Aaron Gutknecht](#)

aaron-gutknecht@gmx.de

Johann Wolfgang Goethe-Universität
Frankfurt a. M., Germany

Target Authors

[Holk Cruse](#)

holk.cruse@uni-bielefeld.de

Universität Bielefeld
Bielefeld, Germany

[Malte Schilling](#)

malte.schilling@uni-bielefeld.de

Universität Bielefeld
Bielefeld, Germany

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

proach is best suited to answer, and what kind of problems it will likely face.

The first of the two main ideas is central to the research area of bio-robotics. If we are able to create an artificial system that exhibits the phenomena of interest, we should be a great deal closer to understanding how these phenomena come about in nature. In order for this approach to lead to valid conclusions, however, the process of reconstruction has to do justice to the systems we are seeking to understand. In the present context we are concerned, above all, with humans and other animals. This means that the way the artificial system achieves the desired results has to be biologically plausible, i.e., it has to be reasonable to assume that the capacities of the organism that we are trying to understand are really based on similar mechanisms. In this vein, [Cruse & Schilling \(this collection\)](#) are realising the basic reactive modules of their system in form of artificial neural networks that were inspired by biological research on, for instance, stick insects (Walknet) and desert ants (Navinet).

The second of the basic ideas derives its plausibility from an evolutionary perspective on psychological faculties. Emotion, attention, the ability to plan future actions, and any other “higher-level” capacities, including consciousness, did not arise suddenly from one generation to the next and independently of pre-existing, more fundamental abilities, such as the ability to control one’s own body and respond adaptively to environmental stimuli. Rather these latter abilities and the interactions between the mechanisms responsible for them might well have been crucial for mental properties to evolve. From this perspective, the idea of reconstructing the evolutionary process by starting with basic reactive structures and examining how through the interaction of these structures unexpected properties might *emerge* seems very promising. Since humans also gradually evolved from simpler organisms, it is natural to assume that the same dependence between reactive structures and “higher-level” phenomena is present in our case as well. The investigation of this dependence might thus provide new insights into the mechanisms underlying human psychology.

But what does it mean exactly to say that a property *emerges* from basic structures? What is an emergent property? The philosophical controversies surrounding the concept of emergence date back over a hundred years and although usage of the term has become increasingly popular in recent years, among both philosophers and scientist, it can hardly be said to have one universal definition. Rather, there are numerous and varied interpretations, a fact which inevitably leads to confusion and misunderstanding (for a good overview see [O’Connor & Wong 2012](#)). It is thus vital to identify precisely what is meant by emergence in any particular case. Notwithstanding this inherent ambiguity, there seems, however, to be a shared idea behind much talk of emergent properties: this is the idea that as systems become increasingly complex they tend to exhibit certain higher-level properties, which are novel or unexpected given their simpler, lower-level, components.

Depending on how this claim is interpreted it can have more or less serious implications regarding the fundamental structure of nature, as well as the structure of science. In order to obtain a particularly strong and at the same time highly influential reading, it must be understood in a two-fold sense. First, as meaning that these properties cannot *even in principle* be predicted or explained on the basis of the lower-level properties of the system and, second, as indicating that such properties are associated with genuinely *new causal powers*, i.e., they make a real difference to the run of events and are not mere epiphenomena (for discussion see [Kim 1999, 2006](#)).¹ This kind of emergence could be called *strong emergence*.² Central to this conception is that emergent properties causally influence the simpler entities from whose organisation they emerge. This sort of causal influence is called “downward causation”, as emergent properties are conceived as

¹ Such conceptions go back to thinkers such as Samuel Alexander, C. L. Morgan, and C. D. Broad, prominent figures in a philosophical movement, which came to be known as “British emergentism”. The following discussion is, however, intended to illustrate the problematic nature of the concept of emergence and not to offer an analysis of the ideas of a particular philosophical school.

² It should be noted that there is no universal definition of the term “strong emergence” in the current literature (for some alternative characterisations see [Chalmers 2006](#); [Bedau 1997](#); [Yates 2013](#)).

higher-level properties arising from certain lower-level properties and relations. Typically, it is assumed that what we find at the lowest level of this hierarchy are the properties and relations of fundamental physical particles. Given this assumption, the existence of emergent properties would entail that a complete description of the fundamental physical organisation of a system might still leave something out. The system might still have some properties that could not be predicted on the basis of such a description and could not be explained in terms of the organisation of its basic physical constituents. Moreover, because emergent properties are causally efficacious, knowledge of the basic physical components of a system and their behaviour may not be sufficient to predict the future evolution of the system. These considerations seem to lead to the conclusion that the meta-scientific thesis, according to which all phenomena can ultimately be explained by the fundamental laws of physics, would turn out to be false. If certain properties belonging to the domains of psychology, biology, or chemistry were emergent properties, these could not even in principle be captured by basic physics alone. All sciences dealing with genuinely emergent properties would remain completely autonomous, positing their own independent laws and explanations. Furthermore, since emergent properties have the ability to causally influence lower-level entities, the fundamental laws of physics would not even suffice to explain processes taking place at the *physical* level (see also [Chalmers 2006](#)).

These are substantial conclusions that could be met with some scepticism. They are also one of the reasons for the fierce controversy surrounding the concept of emergence. Furthermore, the condition that emergent properties are themselves causally efficacious and the general idea of “downward causation” leads to problems in and of itself. This is because there has to be a systematic relationship between emergent and lower-level properties, even though they are conceived as being distinct from another. Often this is expressed by saying that emergent properties are completely determined by lower-level properties and require

them for their existence. In other words, if all lower-level properties of a system are fixed, its emergent properties are also fixed; and without any appropriate lower-level properties, a system cannot have emergent properties. If this weren't the case, it would be unclear in what sense emergent properties *emerge from* lower-level ones ([Kim 2006](#)). If their relationship were completely coincidental, this would surely be an inappropriate description.

Based on this requirement, [Kim \(1999, 2006\)](#) has put forth an influential argument that the idea of “downward causation” is untenable. In summary, Kim's basic argument is this: suppose an emergent property (let's say a feeling of thirst) causes a lower-level property (e.g., a certain activation pattern N in the brain). If feeling thirsty is an emergent property, there have to be appropriate lower-level properties from which it emerges. Let's call these the “emergence base” of feeling thirsty. Now, that feeling thirsty causes N means that there is a natural law that occurrences of feeling thirsty are always followed by occurrences of N (feeling thirsty is nomologically sufficient for N). But since occurrences of feeling thirsty are always accompanied by occurrences of its emergence base, it must also be true that occurrences of its emergence base are followed by occurrences of N. Therefore, if feeling thirsty causes N, its emergence base also causes N. But this makes feeling thirsty completely redundant as a cause of N. Its emergence base is completely sufficient to explain N's occurrence, leaving the feeling of thirst as a mere epiphenomenon. Since this example can easily be generalised, one can conclude that there are no genuine cases of downward causation and hence no genuine emergent properties of the type presently under consideration.

In summary, it can be stated that emergence is a highly controversial concept—not only because of its inherent ambiguity, but also on account of certain varieties of emergentism that have substantial metaphysical and meta-scientific implications as well as a commitment to the problematic idea of downward causation. The crucial questions remaining now are whether [Cruse & Schilling \(this collection\)](#)

provide a clear interpretation of the concept of emergence and whether it provokes the kind of controversy and criticism outlined above. What kind of emergence is involved in their claim that mental states might be construed as emergent properties? In fact, they provide two slightly different characterisations. According to the first, an emergent property is to be understood as a property of a whole system that cannot, *at first sight*, be traced back to the interactions of the systems components. Alternatively, one might say that we cannot, at first sight, predict the emergent properties of a complex system based on our knowledge of its parts and their interaction. Thus, we might be genuinely surprised that the system in question exhibits such properties. Emergence in this sense is sometimes called *weak emergence* (Chalmers 2006). If this is all that it means for a system to have emergent properties, few would raise serious objections. This sort of emergence is just a consequence of our limited knowledge and cognitive capacities and is relative to the judging subject: what might not be immediately predictable for one person might be just so for another. Emergentism, in this sense, has no far-reaching metaphysical or meta-scientific implications and is not committed to any sort of “downward causation”.

Cruse & Schilling (this collection) provide a second, and equally unproblematic, definition of emergence that is specifically tailored for application in the context of robotics. According to that definition, a property of an artificially constructed system is emergent if it was not explicitly implemented by its designers. We might call this *implementational emergence*. This appears to be relatively independent of the sort of “weak” emergence I’ve just described. Even a property not explicitly implemented might be predictable without too much effort, whereas a property deliberately implemented might not be predictable, at least by persons lacking experience or competence. I think that most of the emergent properties Cruse & Schilling (this collection) attribute to their artificial system, reaCog, match both characterisations: they were neither explicitly implemented nor would we immediately expect or predict that reaCog would

exhibit them. At the same time, the properties in question are highly interesting and are not simply insignificant side effects. This is important since, according to the definitions provided by Cruse & Schilling, the claim that an artificial system exhibits emergent properties is, *in and of itself*, not particularly notable. But this depends entirely on what the emergent properties in question precisely are. The finding that reaCog exhibits, in this way, aspects of psychological characteristics, such as emotion or attention and the ability to perform non-trivial body movements, are most certainly of considerable scientific significance. In conclusion, we may say that although the kind of emergentism advocated by Cruse & Schilling does not have the same far-reaching implications as the particularly demanding conception outlined above, it is nonetheless useful and philosophically interesting. This is because it functions as the basis of an intriguing approach to the study of psychological properties, which I shall now endeavour to describe.

Combining the idea of emergence with the idea, outlined above, that in order to understand a system and its properties, it has to be reinvented or reconstructed, we arrive at a fascinating research strategy. The first step consists in observing the behaviour of animals that, although lacking many of the sophisticated abilities with which humans are endowed, are nonetheless capable of flexibly controlling their bodies in order to cope with an unpredictable environment (such as stick insects, desert ants, and honey bees). Based on these observations one then develops a neural network model (e.g., Walknet or Navinet) designed to produce the behaviour observed in the first step. Next, this model is realised in an artificial system (either virtual or robotic) in order to examine to what extent the behaviour produced by the model matches the behaviour of the biological organism on which it is based. If it resembles it to a great extent, this can be taken as *prima facie* evidence that the mechanisms underlying the behaviour are the same for the animal and the robot. Different modules that are constructed in this way are then integrated into a holistic system. Further modules might be added step-by-

step (e.g., Body Model, Attention-Controller, Word-Nets). The result is a complex system (in the present case “reaCog”) the behaviour and properties of which cannot be easily predicted even by its very own designers. The last, and most important step consists of examining whether the system shows characteristics that were not explicitly implemented but instead arise from the dynamic interactions of the system’s components. The most intriguing question in this context is, of course, whether the final system shows aspects of those phenomena that are constitutive of *having a mind*.

Although this is only a rough sketch of the methodology entertained by Cruse & Schilling (this collection), I hope I have captured the essential points sufficiently to proceed with an evaluation of its scope and the possible problems it might face. What kind of questions is the bottom-up approach best suited to answer? Which phenomena or processes can be addressed by research based on this approach? What considerations have to be taken into account in order for the presented research strategy to be successful? Are there any general constraints bio-robotic bottom-up explanations have to meet? As we shall see, the answers to these last two questions are directly connected to two characteristics of the research strategy outlined in the previous paragraph: first, that it involves, at two points, a comparison of the behaviour of significantly different systems and, second, that it is specifically designed to discover emergent properties.

2 The bottom-up approach: Objectives, benefits and constraints

2.1 Mechanisms and the evolution of the mind

The most important aspect of the proposed approach is that it helps to elucidate the *mechanisms* underlying various mental properties. This is possible because many of the basic features of the control system reaCog are known. Using the words of Cruse & Schilling (this collection), it constitutes a “quantitatively defined system”. As all components are realised as artificial

neural networks, all information about the number of neurons, the connection weights between them, and the way individual neurons process information is available. More importantly, however, the basic functional architecture of the system is well understood. Which modules are connected in which ways to other modules, how they receive their input, and what other parts of the system might be affected by their outputs does not have to be figured out by painstaking investigation—as is the case in biological research. Because these facts about reaCog are known, it is possible to provide detailed mechanism descriptions. In this way, reaCog’s ability to plan its future actions by internal simulation can be explained by reference to the interaction of its various sub-modules: a problem detector is activated when sensory input indicates that current behaviour will lead, if continued, to adverse effects for the system (e.g., falling over). This leads to the abortion of current behaviour and activation in the Spreading Activation Layer, which randomly excites the Winner-Takes-All network (WTA-net). After some time, the WTA-net adopts a relaxed state in which only one of its units is active. This active unit in turn stimulates its counterpart in the Motivation Unit Network, leading to activity of the corresponding reactive procedures. These provide motor output that can be redirected to the body model, which then simulates the execution of the proposed behaviour and predicts its likely consequences. If the system predicts that the problem will persist, the process of internal simulation goes on until a solution is found, which can then be used to control the actual movements of the system.

Explanations like these contain a lot of information about which functional subparts of a system are engaged during the exercise of the ability in question. In this particular case it makes clear how the ability to plan ahead, a cognitive ability, depends heavily on basic reactive structures that are designed to control specific leg movements as well as an internal model of the body. The same is true for various other capacities like attention and Theory of Mind. Thus, new insights into the mechanisms responsible for those phenomena in humans could

be gained by considering how body models and motor control mechanisms are realised in our case and how these systems interact. In other words, the bottom-up approach may lead to new directions for future research concerning human psychology by suggesting how specific functional modules interact in order to bring about a particular target phenomenon. Whether this approach is tenable depends on the degree to which findings pertaining to the artificial system might legitimately be used to draw conclusions about human beings. I will propose a number of constraints to ensure that this condition is fulfilled below.

Another class of questions that a bottom-up strategy is well designed to answer has to do with the evolution of cognitive capacities: how did cognitive systems evolve from purely reactive systems? How did emotions, attention, or even consciousness arise? What are the natural precursors of these phenomena? Cruse & Schilling (this collection) show convincingly that no completely new neural modules are needed in order for such properties to occur. Rather, minor changes in the basic architecture might suffice to generate radical extensions of the abilities of a system. In this way, a reactive system with a body model can acquire the ability to plan ahead if it is able to disconnect its motor system from the physical body and instead send the motor signals to its internal body model. No novel “planning module” is needed. Already existing modules just have to become dissociable and can thus acquire new functions (Cruse 2003). In addition, the target paper suggests an answer to the question of the evolutionary function of cognition understood as the ability to plan ahead: it was the necessity of being able to control a complex body in a complex environment that made this ability highly valuable. Detecting problems by perception, finding innovative solutions by internal simulation and acting on them are capacities that are extremely advantageous for any organism possessing a body with a high number of redundant degrees of freedom (see Cruse 2003). This is in line with, and actually extends, the widespread assumption that the evolutionary function of cognition is to deal with environmental complexity (Godfrey-Smith 2002).

2.2 Constraints on bio-robotic bottom-up explanations

In the previous paragraph we saw that the framework Cruse & Schilling (this collection) present is well-equipped to give new insights into the underlying mechanisms of psychological phenomena and the evolution of cognition, as well as a promising approach to creating highly flexible and intelligent robots. There are, however, some problems the proposed strategy has to face, especially if the control structures become increasingly complex. I therefore want to suggest a set of three constraints on good bottom-up explanations of biological/psychological phenomena.

1. *Adequate matching criteria*.³ At two points the research strategy described in section 1 involves a comparison between the behaviour of an artificial system on the one hand and a biological system on the other. First, this is the case in the development of neural network models of animal behaviour. In this context, the comparison is used to ascertain whether the proposed model of the mechanisms underlying certain capacities (e.g., walking) really reproduces the original behaviour of the animal (e.g., a stick insect). Second, there is a similar process of comparison involved in the application of psychological concepts to the complete system. At different points in their discussion, Cruse & Schilling (this collection) argue that their system has certain mental capacities because it exhibits behaviour (or would exhibit it if certain extensions were implemented) connected to those mental capacities in humans. So, for example, just as the performance of athletes might worsen if they consciously attend to what they are doing, the activation of the attention controller in *reaCog* can lead to poorer results compared to unimpeded execution of the reactive procedures.

Both processes of comparison require criteria to identify when the behaviour of the artificial system and that of the biological system

3 I credit this term to Datteri & Tamburrini 2007.

are relevantly similar, i.e., similar enough in order to provide evidence for the claim that similar mechanisms are at work in both cases or that the artificial system and the biological system share certain psychological characteristics (Datteri & Tamburrini 2007). The difficulty of finding such criteria increases the more the bodies of the compared systems differ. In some cases they might nonetheless be easy to find and relatively uncontroversial. This, however, is not always the case. For instance, in their discussion of emotions—and more specifically the emotion of happiness—Cruse & Schilling (this collection) suggest that by increasing the threshold of the problem detector reaCog would take more risks, thus behaving similarly to humans when they are happy. Now, the question is whether the kind of risky behaviour exhibited by reaCog when the threshold of its problem detector is increased is the same kind of risky behaviour humans exhibit when they are happy. Only if this condition is fulfilled can the similarity be taken as evidence that reaCog shows aspects of the emotion of happiness.

2. *Biological plausibility*: Any proposed mechanism should be biologically plausible, i.e., it has to be reasonable to assume that the capacities of the organism that we are trying to understand are really based on such a mechanism. This can, at least to some degree, be ensured by trying to create similarities between the artificial and the biological organism on a basic structural level, for example by using artificial neural networks. Furthermore, it is necessary to decide how fine-grained the model should be. Should the model take brain structures, neurons, or sub-cellular elements as its basic building blocks? Should intracellular processes be neglected or are they important? The answer will of course always be relative to our particular epistemic goals. Finally, there are different options regarding the way artificial neurons process information, i.e., how they calculate their output value depending on the weighted sum of their inputs. All these factors might turn out to be important if the results are to

be used to infer biological mechanisms. The requirement of biological plausibility shouldn't, however, be overemphasised. Cruse & Schilling (this collection) stress that they are not trying to present a realistic model of neuronal activity in living organisms. Hence, they are using biologically implausible, non-spiking artificial neurons as the basic elements of their architecture, while noting that some authors (referring to Singer 1995) have located the neural basis of consciousness in synchronously oscillating spikes. This, however, is not a weighty objection to the proposed approach since it is designed as a *functional approach*. The question is: how do different functional subsystems like a system for controlling the swing-movement of a leg, a system modelling the robot's body, and a system allowing for the selection of different internal states interact in order to produce certain emergent phenomena? Therefore, the concrete physical realisation of these subsystems is of only secondary importance.

3. *Transparency*:⁴ Doubts about the strategy of using artificial systems in order to understand biological systems arise because even if we were to create an extremely intelligent robot, it would not necessarily help us to understand the mechanisms underlying its intelligence. Rather, we might be faced with yet another complex system whose workings we do not understand (Holland & Goodman 2003). Now, the approach Cruse & Schilling (this collection) present is specifically designed to discover emergent properties, i.e., properties that were not explicitly implemented. This means that there will be a high risk of finding properties in the complete system that cannot be readily provided with a clear-cut mechanistic explanation involving the co-operation of the system's components. Although the explanations of the occurrence of various psychological properties presented in the present paper are quite convincing, the

4 The concept of transparency has a number of other well-established interpretations in the literature that should not be confused with the one at issue in the present context. These include, for example, "semantic" (Clark 1989) and "phenomenal" (Metzinger 2004) transparency.

bottom-up strategy might eventually exhaust its potential if the complexity of the system is further increased.

3 Future perspectives: The social insect

I would like to conclude by briefly proposing a perspective for future research based on the system *reaCog*. As presented, its ability to interact and cooperate with other agents is fairly restricted. At the same time, the pre-requisites of a broader social extension of the system seem to be in place. The present paper already shows how *reaCog* could be equipped with the capacities to recognize the behaviour of others and apply a Theory of Mind. In their 2011 paper, Cruse & Schilling further propose that by implementing a two-body model (a “We-model”) *reaCog* might be capable of cooperative behaviour using shared goals. Integration and further expansion of such social capacities, and their application in an actual robot, seems promising considering the importance of social interaction in processes such as language acquisition and emotional regulation. Some have even suggested that the presence of other agents in the environment, or, in other words, dealing with social complexity, was a dominant factor in the evolution of sophisticated cognitive abilities (Humphrey 1976). Thus bio-robotic research in this direction might provide new insights into the mechanisms underlying such developmental and evolutionary processes. Moreover, a social extension of *reaCog* might eventually shed light on potential emergent phenomena *on a group level*, such as labour division, collective planning, social hierarchies and, most fundamentally, joint action coordination. What high-level social phenomena emerge when multiple bio-robotic systems like *reaCog* interact with each other?

Cruse and Schilling’s system seems particularly well-suited to further illuminate motor theories of social cognition. According to such theories, the important social cognitive capacity of understanding another’s actions is directly linked to mechanisms that are active when the observer performs similar actions (Gallese et al. 2004; for criticism see Jacob & Jeannerod

2005). The underlying neural mechanism has come to be known as the mirror-neuron system. Furthermore, there is evidence that the mirror-neuron system plays a role in certain aspects of self-consciousness. For instance, Uddin (2007; see also Molnar-Szakacs & Uddin 2013) suggests that this is the case for representations of the physical self, and ascribes frontoparietal mirror-neuron areas an important function for self-recognition (especially the recognition of one’s own face). As mirroring mechanisms can be integrated in *reaCog* as well, this opens the possibility of further investigating motor theories of social cognition and the relation between internal motor simulation and the self in a quantitatively defined system.

An ability that is highly important for human social interaction is the ability to communicate using language. At this point, the linguistic capacities of *reaCog* still seem quite inflexible and limited in scope. A highly interesting extension of this system would be to provide it with the means to learn words and their meanings by interaction with other agents. Some of the pre-requisites, like the ability to internally simulate the behaviour of others, could, as Cruse and Schilling argue, be implemented in *reaCog* by using its internal body-model to represent another agent. Robotic research in this direction was performed by Steels & Spranger (2009). Their artificial systems are capable of autonomously acquiring a simple language consisting of words for specific body postures. After learning is complete, the artificial agents are able to reliably assume body postures on verbal command by other agents. Since social learning has also been implicated in the process of concept formation (Steels 2002), the proposed extension might also foster our understanding of this intriguing phenomenon.

4 Conclusion

In conclusion it can be stated that Cruse & Schilling (this collection) present a highly fascinating research strategy that is well worth pursuing. The bottom-up approach can provide us with new insights regarding the functional mechanisms underlying psychological phenom-

ena and their evolution. Although the notion of emergence is central to it, Cruse & Schilling ([this collection](#)) avoid the philosophical controversies surrounding this concept by interpreting it in a less demanding, yet interesting and useful way. There are, however, a number of constraints that explanations based on the bottom-up approach have to meet. First, since Cruse & Schilling's ([this collection](#)) strategy involves, at two points, a comparison between markedly different systems, criteria are needed according to which we can determine whether the two systems exhibit relevantly similar behaviour. Second, the structural architecture of the artificial system must have an adequate degree of biological plausibility. And finally, it has to be ensured that increasing the complexity of the system does not lead to the practical impossibility of elucidating the mechanisms underlying its emergent properties.

A promising next step for bottom-up research as presented by Cruse & Schilling ([this collection](#)) would be to take it to the level of social interaction. An extensive social extension of their system could shed light on a wide range of intriguing phenomena. Is it possible to discover emergent properties on a group level? In what precise way are mirroring mechanisms involved in social cognition? What role do such mechanisms play for the phenomenon of self-consciousness? What role do reactive structures and internal body-models play in the processes of language acquisition and comprehension? Of course this is only a small selection of the questions further bio-robotic research might contribute to answering. Cruse & Schilling ([this collection](#)) made clear that starting from the bottom is a strategy with enormous scientific significance. There is no doubt that this work will make an important contribution to a plethora of research projects in the future.

References

- Bedau, M. A. (1997). Weak emergence. *Philosophical Perspectives*, 11 (s11), 375-399.
[10.1111/0029-4624.31.s11.17](#)
- Chalmers, D. J. (2006). Strong and weak emergence. In P. Davies & P. Clayton (Ed.) *The re-emergence of emergence* (pp. 244-256). Oxford, UK: Oxford University Press.
- Clark, A. (1989). *Microcognition: Philosophy, cognitive science, and parallel distributed processing*. Cambridge, MA: MIT Press.
- Cruse, H. (2003). The evolution of cognition - a hypothesis. *Cognitive Science*, 27 (1), 135-155.
[10.1207/s15516709cog2701_5](#)
- Cruse, H. & Schilling, M. (2011). From egocentric systems to systems allowing for theory of mind and mutualism. In T. Lenaerts, M. Giacobini, H. Bersini, P. Bourguin, M. Dorigo & R. Doursat (Ed.) *Advances in artificial life, ECAL 2011. Proceedings of the eleventh european conference on the synthesis and simulation of living systems* (pp. 185-192). Cambridge, MA: MIT Press.
- Clark, A. & Schilling M. (2015). Mental states as emergent properties. In T. Metzinger & J. M. Windt (Ed.) *Open MIND* (pp. 1-39). Frankfurt a. M., GER: MIND Group.
- Datteri, E. & Tamburrini, G. (2007). Biorobotic experiments for the discovery of biological mechanisms. *Philosophy of Science*, 74 (3), 409-430.
[10.1073/pnas.1015390108](#)
- Gallese, V., Keysers, C. & Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends in Cognitive Sciences*, 8 (9), 396-403.
[10.1016/j.tics.2004.07.002](#)
- Godfrey-Smith, P. (2002). Environmental complexity and the evolution of cognition. In R. Sternberg & J. Kaufman (Ed.) *The evolution of intelligence* (pp. 233-249). Hove, UK: Psychology Press.
- Holland, O. & Goodman, R. (2003). Robots with internal models a route to machine consciousness? *Journal of Consciousness Studies*, 10 (4-5), 77-109.
- Humphrey, N. K. (1976). The social function of intellect. In P. P. G. Bateson & R. A. Hinde (Ed.) *Growing point in ethology* (pp. 303-317). Cambridge, UK: Cambridge University Press.
- Jacob, P. & Jeannerod, M. (2005). The motor theory of social cognition: A critique. *Trends in Cognitive Sciences*, 9 (1), 21-25.

- Kim, J. (1999). Making sense of emergence. *Philosophical studies*, 95 (1), 3-36. [10.1023/A:1004563122154](https://doi.org/10.1023/A:1004563122154)
- (2006). Emergence: Core ideas and issues. *Synthese*, 151 (3), 547-559. [10.1093/acprof:oso/9780199585878.001.0001](https://doi.org/10.1093/acprof:oso/9780199585878.001.0001)
- Metzinger, T. (2003). *Being no one: The self-model theory of subjectivity*. MIT Press.
- Molnar-Szakacs, I. & Uddin, L. Q. (2013). The emergent self: How distributed neural networks support self-representation. *Handbook of neurosociology* (pp. 167-182). Dordrecht, NL: Springer.
- O'Connor, T. & Wong, H. Y. (2012). Emergent properties. *Stanford Encyclopedia of Philosophy*. <http://plato.stanford.edu/entries/properties-emergent/>
- Singer, W. & Gray, C. M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annual review of neuroscience*, 18 (1), 555-586. [10.1146/annurev.ne.18.030195.003011](https://doi.org/10.1146/annurev.ne.18.030195.003011)
- Steels, L. & Kaplan, F. (2002). AIBO's first words: The social learning of language and meaning. *Evolution of Communication*, 4 (1), 3-32. [10.1075/eoc.4.1.03ste](https://doi.org/10.1075/eoc.4.1.03ste)
- Steels, L. & Spranger, M. (2009). How experience of the body shapes language about space. In M. Kaufmann (Ed.) *IJCAI'09: Proceedings of the 21st international joint conference on Artificial intelligence*. San Francisco, CA: Morgan Kaufmann.
- Uddin, L. Q., Iacoboni, M., Lange, C. & Keenan, J. P. (2007). The self and social cognition: The role of cortical midline structures and mirror neurons. *Trends in Cognitive Sciences*, 11 (4), 153-157. [10.1016/j.tics.2007.01.001](https://doi.org/10.1016/j.tics.2007.01.001)
- Yates, D. (2013). Emergence. In H. Pashler (Ed.) *Encyclopedia of the Mind* (pp. 283-287). San Diego, CA: SAGE Reference.

The Bottom-Up Approach: Benefits and Limits

A Reply to Aaron Gutknecht

Holk Cruse & Malte Schilling

Aaron Gutknecht supports our bottom-up approach, specifies possible limits and highlights interesting future aspects. His added perspective is valuable and interesting to us. As we fully agree with his view, we only add some complementary remarks.

Keywords

Bottom-up approach | Concept clarifying machine | Emergence | Emotions in arthropods | We-model

Authors

[Holk Cruse](#)

holk.cruse@uni-bielefeld.de

Universität Bielefeld

Bielefeld, Germany

[Malte Schilling](#)

malte.schilling@uni-bielefeld.de

Universität Bielefeld

Bielefeld, Germany

Commentator

[Aaron Gutknecht](#)

aaron-gutknecht@gmx.de

Johann Wolfgang Goethe-Universität

Frankfurt a. M., Germany

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität

Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University

Melbourne, Australia

1 Introduction

We appreciate the comments given by Aaron Gutknecht very much, in particular his discussion and clarification of the term “emergence” and its philosophical background. This discussion comprises a sensible completion of our article going beyond the scope of our expertise. In this context, Aaron Gutknecht correctly states that our way of

using the term “emergence” may cover two aspects, one called “weak emergence”, the other he addressed as “implementational emergence”. We have – possibly forming some kind of common denominator – a third characterization in mind, one that covers different description levels: a phenomenon is considered emergent if it turns out

that known properties of the network could also be characterized on a different level of description than the one currently used. On this different level the phenomenon conceptually constitutes a term or definition. If we, for example, describe the structure and function of reaCog on the neuronal level, we may realize at some point that there are behavioral aspects which could, by an outside observer, be characterized by a term that is not defined at a neuronal level of description, such as, for example, “intention”.

2 The bottom-up approach

This way of using the term emergence is directly related to the bottom-up approach applied here. This approach is inspired by Feynman, who stated that we understand a system only when we are able to construct it (in [Hawking 2001](#)) and may be even dated back to [Giambattista Vico \(1710\)](#). The bottom-up approach allows us to study the extent to which linguistic concepts proposed in the literature may correspond to properties realized by our artificial system. If one was not prepared to accept that a specific concept would correspond to selected properties of the artificial system, either the linguistic concepts might be adapted accordingly, or the artificial system might be judged as to show deficits. The latter case could then give rise to adapt the current simulation model to better match the verbal proposal given. This capability of the bottom-up approach led [Manuela Lenzen \(2014\)](#) to characterize reaCog as a “concept clarifying machine” (“Begriffspräzisierungsmaschine”).

3 Possible limits of the bottom-up approach

Aaron Gutknecht further proposes a well-chosen list of issues that should be taken into account when following a bottom-up approach as proposed here, namely “adequate matching criteria”, “biological plausibility” and “transparency”.

Concerning the first issue, “adequate matching criteria”, Aaron Gutknecht addresses a possibly critical point. In section 8 (*Emotions*), we characterize happiness by the property that risky decisions are made more probable. We admit that

our example is formulated in a sketchy way, only addressing one basic aspect for illustration. There are, however, more deeply founded examples that have been briefly referred to in the main text and will be explained in more detail here. Two recent studies, one in crayfish, the other in the fruitfly, provide strong hints that emotion-like states can be found in simple organisms as arthropods or, more specifically in the latter-case, insects. In crayfish, [Fossat et al. \(2014\)](#) have convincingly shown that context-independent, anxiety-like behavior can be induced by experimentally applied stress or by application of serotonin. Both methods lead the animals to avoid illuminated sections of their environment which they are normally interested to explore. Anxiety is related to fear but considered a secondary emotion that occurs after the stressing signal has disappeared. Thus, the probability of selecting specific behaviors, in this case exploration of illuminated places, is decreased. This avoidance behavior could be abolished after application of drugs that are known to have anxiolytic effects in mammals. Applied to reaCog, these results could be interpreted in the following way. Emotion-like states would not only influence the global WTA net, but also thresholds of local, lower level WTA networks that are responsible for switching between different procedures.

Another interesting case has been reported by [Yang et al. \(2014\)](#) in *Drosophila*. These animals learnt that various behaviours selected in trying to avoid a problem, in this case escape from a heated ground, were not successful. As a consequence, they ended up in a state of passivity. This result has been discussed as an example of “learned helplessness”, which is considered an animal model of depression. In our framework, this could simply be realized by freezing activity in the Spreading Activation Layer network that provides input to the WTA net (section 4).

Concerning the second issue, “biological plausibility”, we fully support Gutknecht’s perspective and have only a minor aside. Application of non-spiking neurons is not necessarily biologically implausible. Rather, non-spiking neurons do exist in invertebrate and in vertebrate brains. They play important functional roles, but are generally less well-known, mainly

because investigating them involves methodological problems that are more difficult than those of spiking neurons. The third issue, “transparency”, addresses the view that the bottom-up strategy may eventually exhaust its potential when the complexity of the system is further increased. Although we agree with Gutknecht here, we would like to add that the bottom-up approach still bears the advantage that, as the details of such a system are known, its properties can be thoroughly analyzed by physical and/or mathematical methods. This ability, of course, does not guarantee that one will find answers in such a hypothetical case, but there are various methods available to address such questions. Further, we believe that the problem of lacking transparency may not happen to occur too often. This belief is supported by the observation that already our simple system, *reaCog*, appears to be able to reach integration levels characterized by terms such as intention, volition and consciousness.

4 What should be done next?

Aaron Gutknecht closes his comments by considering future aspects. Again, we agree with his recommendations and have, partly, indeed started with two of the aspects addressed. We applied the internal model in a cooperative scenario in which the visual impression of another agent performing an action was mapped onto the system’s own internal body model. In this way the internal model was driven by the visual input and the internal model reenacted what the other agent was doing. This mapping allows one to connect the experiences of somebody else to one’s own action repertoire as one steps into the shoes of the other (Schilling 2011; see also Gallese & Cuccio this collection). Second, as mentioned in the main text, shared circuits are required for an agent to represent the action of a partner (Cruse & Schilling this collection, figure 9). In order to allow for ToM, an additional separate representation of the partner’s memory is required (figure 10). To be able to apply a supermodel (or we-model, Tomasello 2009), a more complex model is required (see Cruse & Schilling 2011, figure 6).

5 Conclusion

The bottom-up approach advocated here to understand higher-level phenomena may be considered a non-Platonic approach that aims to construct artificial, but strongly biologically inspired systems. These systems should be able to simulate complex behavioral tasks, but do so by application of simple elements, artificial neurons, and a simple decentralized neuronal architecture. If successful one could then study whether more abstract concepts introduced in psychology or philosophy, for example, could sensibly be applied to such a system. We claim to have shown an example supporting this approach.

References

- Cruse, H. & Schilling, M. (2011). From egocentric systems to systems allowing for theory of mind and mutualism. In R. Doursat (Ed.) *Proceedings of the ECAL 2011, Paris* (pp. 731-738). Cambridge, MA: MIT Press.
- (2015). Mental states as emergent properties. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Fossat, P., Bacqué-Cazenave, J., De Deurwaerdère, P., Delbecq, J.-P. & Cattaert, D. (2014). Anxiety-like behavior in crayfish is controlled by serotonin. *Science*, 344 (6189), 1293-1297. [10.1126/science.1248811](https://doi.org/10.1126/science.1248811)
- Gallese, V. & Cuccio, V. (2015). The paradigmatic body. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1-23). Frankfurt a. M., GER: MIND Group.
- Hawking, S. (2001). *The universe in a nutshell*. London, UK: Bantam Press.
- Lenzen, M. (2014). Der sensible Hector - Interaktion mit Robotern. *Frankfurter Allgemeine Zeitung* (2014.9.2014, p.N3)
- Schilling, M. (2011). Learning by seeing - Associative learning of visual features through mental simulation of observed action. In R. Doursat (Ed.) *Proc. of the ECAL 2011, Paris* (pp. 731-738). Cambridge, MA: MIT Press.
- Tomasello, M. (2009). *Why we cooperate*. Cambridge, MA: MIT Press.
- Vico, G. (1710). De antiquissima italorum sapientia. In R. Parenti (Ed.) *Opere*. Naples, I: F. Rossi.
- Yang, Z., Bertolucci, F., Wolf, R. & Heisenberg, M. (2014). Flies cope with uncontrollable stress by learned helplessness. *Current Biology*, 23 (9), 799-803. [10.1016/j.cub.2013.03.054](https://doi.org/10.1016/j.cub.2013.03.054)

Why and How Does Consciousness Seem the Way it Seems?

Daniel C. Dennett

Are-expression of some of the troublesome features of my oft-caricatured theory of consciousness, with new emphases, brings out the strengths of the view and shows how it comports with and anticipates the recent introduction of Bayesian approaches to cognitive science.

Keywords

Bayes | Consciousness | Hume | Inversion | Qualia | Transduction

Author

Daniel C. Dennett

daniel.dennett@tufts.edu

Tufts University

Medford, MA, U.S.A.

Commentator

David Baßler

davidhbassler@gmail.com

Johannes Gutenberg-Universität

Mainz, Germany

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität

Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University

Melbourne, Australia

1 Introduction

People are often baffled by my theory of consciousness, which seems to them to be summed up neatly in the paradoxical claim that consciousness is an illusion. How could that be? Whose illusion? And would it not be a *conscious* illusion? What a hopeless view! In a better world, the principle of charity would set in and they would realise that I probably had something rather less daft in mind, but life is short, and we'll have one less difficult and counterintuitive theory to worry about if we just dismiss Dennett's as the swiftly self-refuting claim that consciousness is an illusion. Other theorists, including, notably, [Nicholas Humphrey \(2006, 2011\)](#), [Thomas Metzinger \(2003, 2009\)](#)

and [Jesse Prinz \(2012\)](#), know better, and offer theories that share important features with mine. I toyed with the idea of trying to re-offer my theory in terms that would signal the areas of agreement and disagreement with these welcome allies, but again, life is short, and I have found that task simply too much hard work. So with apologies, I'm going to restate my position with a few new—or at least newly emphasized—wrinkles, and let them tell us where we agree and disagree.

I take one of the usefully wrong landmarks in current thinking about consciousness to be Ned Block's attempt to distinguish “phenomenal consciousness” from “access consciousness.”

His view has several problems that I have pointed out before (Dennett 1994, 1995, 2005; Cohen & Dennett 2011), but my criticisms have not been sufficiently persuasive, so I am going to attempt, yet again, to show why we should abandon this distinction as scientifically insupportable and deeply misleading. My attempt should at least help put my alternative view in a better light, where it can be assayed against the views of Block and others. Here is the outline, couched in terms that will have to be clarified and adjusted as we go along:

1. There is no double transduction in the brain. (section 1)
Therefore there is no second medium, the medium of consciousness or, as I like to call this imaginary phenomenon, the *MEDium*. Therefore, qualia, conceived of as states of this imaginary medium, do not exist.
2. But it seems to us that they do. (section 2)
It seems that qualia are the source or cause of our judgments about phenomenal properties (“access consciousness”), but this is backwards. If they existed, they would have to be the *effects* of those judgments.
3. The seeming alluded to in proposition 2 is to be explained in terms of Bayesian expectations. (section 3)
4. Why do qualia seem simple and ineffable?
This is an effect, a byproduct, an artifact of “access consciousness.” (section 4)
5. *Whose* access? Not a witness in the Cartesian Theater (because there is no such functional place). (section 5)
The access of other people! Our “first-person” subjectivity is shaped by the pressure of “second-persons”—interlocutors—to have practical access to what is going on in our minds.
6. A thought experiment shows how even color qualia can be understood as Bayesian projections.

2 There is no double transduction in the brain

The arrival of photons on the retina is transduced thanks to rhodopsin in the rods and

cones, to yield spike trains in the optic nerve (I’m simplifying, of course). The arrival of pressure waves at the hair cells in the ear are similarly transduced into spike trains in the auditory nerve, heat and pressure are transduced into yet more spike trains by subcutaneous receptors, and the presence of complex molecules in the air we breathe into our noses is transduced by a host of different transducer molecules in the nasal epithelium. The common medium of spike trains in neuronal axons is well understood, but used to be regarded as a baffling puzzle: how could spike trains that were so alike in their physical properties and patterning underlie such “phenomenally” different phenomena as sight, hearing, touch, and smell? (see Dennett 1978, for an exposure of the puzzle.) It is still extremely tempting to imagine that vision is like television, and that those spike trains get transduced “back into subjective color and sound” and so forth, but we know better, don’t we? We don’t have to strike up the little band in the brain to play the music we hear in our minds, and we don’t have to waft molecules through the cortex to be the grounds for our savoring the aroma of bacon or strawberries. There is no second transduction. And if there were, there would have to be a third transduction, back into spike trains, to account for our ability to judge and act on the basis of our subjective experiences. There might have been such triple transductions, and then there would have been a Cartesian Theater Deluxe, like the wonderful control room in the film *Men in Black*. But biology has been thrifty in us: it’s all done through the medium of spike trains in neurons. (I recognize that dualists of various stripes—a genus thought extinct not so many years ago—will want to dig in their heels right here. I will ignore their howls for the time being, thinking that I can dispatch them later in the argument when I provide an answer to their implied question “What else could it be?”)

So there is no *MEDium* into which spike trains are transduced. Spike trains are discriminated, elaborated, processed, reverberated, re-entered, combined, compared, and contrasted—but not transduced into anything else until some of them activate effectors (neuromuscular

junctions, hormone releasers, and the like) which do the physical work of guiding the body through life. The rich and complex interplay between neurons, hundreds of neuromodulators, and hormones is now recognized, thanks to the persuasive work of Damasio and many others, as a central feature of cognition and not just bodily control, and one can speak of these interactions as transduction back and forth between different media (voltage differences and biochemical accumulations, for instance)—but none of these is the imagined *MEdium* of subjective experience.

So there just is no home in the brain for qualia as traditionally conceived. My point can be clarified by a simple comparison between two well-understood media: cinema film and digital media. First imagine showing some stone-age hunter-gatherers a movie using a portable Super-8 film projector. Amazing, they would think, but when they were then shown the frames of film up close, they would readily understand—I daresay—that this was not magic, because there were little blobs of color on each frame. (The soundtrack might still be baffling, but perhaps they would hold the film up to their ears and decide, eventually, that the sounds were just too faint for them to hear with their naked ears.) Then show them a film on a portable DVD player, and demonstrate the powers of the removable, interchangeable disks, and let them ponder the question of how such a disk managed to store all the sounds and colors they just observed on the screen. It would probably be tempting for them to declare that it *must* be magic—dualism, in other words. But with a little instruction, they could no doubt catch on to the idea that you don't have to represent color with color, sound with sound. You can *transduce* color, sound—anything, really—into a system of patterns of differences (0s and 1s, spike trains, ...) and then *transduce* the elements of that system back into color and sound with playback equipment. This could lay magic to rest.

I had better make my implicit claim explicit, at the risk of insulting some readers: if you think there *has* to be a medium in the brain (or in a dualistic mind) in which subjective colors,

sounds, and aromas are *rendered*, you are making the stone-ager mistake. This, I have come to believe, is the stone wall separating my view from wider acceptance. People pay attention to my arguments, and then, confronted with the prospect that qualia, as traditionally conceived, are not needed to explain their subjectivity, they just dismiss the idea as extravagant. “OF COURSE there are qualia!” This thought experiment is meant to shock them: your confidence here, I am saying, is no better grounded than the imagined confidence of the stone-agers that there just *have to be* colors and sounds on the DVD for it to convey colors and sounds to the playback machine. A failure of imagination mistaken for an insight into necessity. “But when I have a tune running through my head, it has pitch and tempo, and the timbre of the instruments is there just as if I were listening to a live performance!” Yes, and for that to be non-magically the case, there has to be a representation of the tune that progresses more or less in real time, and that specifies pitch and timbre, but that can all be accomplished without transduction, without further *rendering*, in the sequence of states of neural excitation in auditory cortex.

Vision isn't television, and audition isn't radio. We are accustomed, now, to playback devices that do transduce the signals back into the colors and sounds from which they were transduced, but we need to take advantage of our twenty-first century sophistication and recognize that the second transduction is optional! The information is in the signal, and all that information can be processed, discriminated, translated, re-coded, simplified, embellished, categorized, tagged, adjusted, and used to guide behavior without ever being transduced back into colors and sounds (or “subjective” colors and sounds).

3 It still seems that qualia exist

But it sure seems that qualia exist, in spite of the foregoing! How could they not? Aren't they needed, for instance, to be the source or cause of our judgments about them? If I have a conviction that I'm seeing an American flag after-

image (see [figure 1](#)), and note that the lowest short red stripe intersects the central cross, doesn't there have to be the red stripe I deem myself to be experiencing? Isn't the presence of that red stripe *somewhere* a necessary condition for me seeming to see a red stripe? No, and the alternative has been at least dimly understood since Hume's brilliant discussion of our experience of causation.

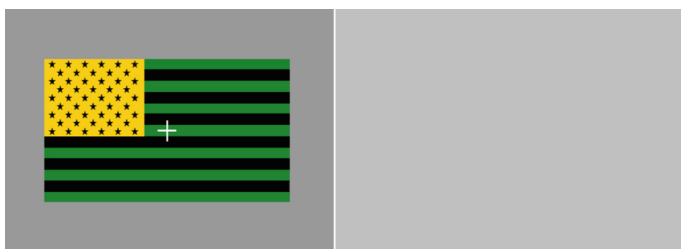


Figure 1: Inverted American Flag.

Consider what I will call Hume's Strange Inversion (cf. [Dennett 2009](#)). We think we see causation because the causation in the world directly causes us to see it—the same way round things in daylight cause us to see round things, and tigers in moonlight cause us to see tigers. When we see the thrown ball causing the window to break, the causation itself is somehow perceptible “out there.” Not so, says [Hume \(1739, section 7 “Of the idea of necessary connexion”\)](#). What causes us to have the idea of causation is not something external but something internal. We have seen many instances of *As followed by Bs*, Hume asserts, and by a process of roughly Pavlovian conditioning (to put it anachronistically) we have been caused by this series of experiences to have in our minds a disposition, when seeing an A, to expect a B—even before the B shows up. When it does, this *felt* disposition to expect a B is mis-identified as an external, *seen* property of causation. We think we experience causation between A and B, when we are actually experiencing our internal judgment “here comes a B” and “projecting” it into the world. This is a special case of the mind's “great propensity to spread itself on external objects” ([Hume 1739, I, xiv](#)). In fact, Hume insisted, what we do is misinterpret an inner “feeling”—an anticipation—as an external property. The “customary transition” in our

minds is the source of our sense of causation, a quality of “perceptions, not of objects,” but we mis-attribute it to the objects, a sort of benign user-illusion, to speak anachronistically again. As Hume notes, “the contrary notion is so riveted in the mind” that it is hard to dislodge. It survives to this day in the typically unexamined assumption that all perceptual representations must be flowing inbound from outside.

Hume wrote that the ‘mind has a great propensity to spread itself on external objects’ (T 1.3.14.25; SBN 167) and that we ‘gild and stain’ natural objects ‘with the colours borrowed from internal sentiment’ (EPM Appendix 1.19; SBN 294). These metaphors have invited a further one: that of ‘projection’ and its cognates. Though not Hume's own, the projection metaphor is now so closely associated with him, both in exegetical and non-exegetical contexts, that the phrase ‘Humean projection’ is something of a cliché in philosophical discourse. ([Kail 2007, p. 20](#))

Here are a few other folk convictions that need Strange Inversions: sweetness is an “intrinsic” property of sugar and honey, which causes us to like them; observed intrinsic sexiness is what causes our lust; it was the funniness out there in the joke that caused us to laugh ([Hurley et al. 2011](#)). There is no more familiar and appealing verb than “project” to describe this effect, but of course everybody knows it is only metaphorical; colors aren't literally projected (as if from a slide projector) out onto the front surfaces of (colorless) objects, any more than the idea of causation is somehow beamed out onto the point of impact between the billiard balls. If we use the shorthand term “projection” here to try to talk, metaphorically, about the mismatch between manifest and scientific image ([Sellars 1962](#)), what is the true long story? What is literally going on in the scientific image? A large part of the answer emerges, I propose, from the predictive coding perspective. Every organism, whether a bacterium or a member of *Homo sapiens*, has a set of things in the world that matter to it and which it (therefore) needs to

discriminate and anticipate as best it can. Call this the ontology of the organism, or the organism's "Umwelt" (von Uexküll 1957). This does not yet have anything to do with consciousness but is rather an "engineering" concept, like the ontology of a bank of elevators in a skyscraper: all the kinds of things and situations the elevators need to distinguish and deal with. An animal's "Umwelt" consists in the first place of affordances (Gibson 1979), things to eat or mate with, openings to walk through or look out of, holes to hide in, things to stand on, and so forth. We may suppose that the "Umwelt" of a starfish or worm or daisy is more like the ontology of the elevator than like our manifest image. What's the difference? What makes our manifest image manifest (to us)?

4 Bayesian expectations

Here is where Bayesian expectations (see Clark 2013) could play an iterated role: our ontology (in the elevator sense) does a close-to-optimal job of representing the things in the world that matter to the behavior our brains have to control (cf. Metzinger 2003, on our world models). Hierarchical Bayesian predictions accomplish this, generating affordances galore: we expect solid objects to have backs that will come into view as we walk around them, doors to open, stairs to afford climbing, cups to hold liquid, etc. But among the things in our Umwelt that matter to our wellbeing are ourselves! We ought to have good Bayesian expectations about what we will do next, what we will think next, and what we will expect next! And we do. Here's an example:

Think of the cuteness of babies. It is not, of course, an "intrinsic" property of babies, though it seems to be. What you "project" out onto the baby is in fact your manifold of "felt" dispositions to cuddle, protect, nurture, kiss, coo over, ... that little cutie-pie. It's not just that when your cuteness detector (based on facial proportions, etc.) fires, you have urges to nurture and protect; you expect to have those very urges, and that manifold of expectations just is the "projection" onto the baby of the property of cuteness. When we expect to see a

baby in the crib, we also expect to "find it cute"—that is, we expect to expect to feel the urge to cuddle it and so forth. When our expectations are fulfilled, the absence of prediction error signals is interpreted as confirmation that, indeed, the thing in the world with which we are interacting has the properties we expected it to have. Without the iterated expectations, cuteness could do its work "subliminally," outside our notice; it could be part of our "elevator ontology" (the ontology that theorists need to posit to account for our various dispositions and talents) but not part of *our* ontology, the things and properties we can ostend, reflect on, report, discuss, or appeal to when explaining our own behavior (to ourselves or others). Cuteness as a property passes the Bayesian test for being an objective structural part of the world we live in (our *manifest* manifest image), and that is all that needs to happen. *Any further "projection" process would be redundant. What it is to experience a baby as cute is to generate the series of expectations and confirmations just described. What is special about properties like sweetness and cuteness is that their perception depends on particularities of the nervous systems that have evolved to make much of them. The same is of course also true of colors. This is what is left of Locke's (and Boyle's) distinction between primary and secondary qualities.*¹

Similarly, when we feel the urge to judge something about "that red stripe" (in the American flag afterimage (see Figure 1) that hovers in our visual field, we have the temptation to insist that there is a red stripe—there has to be!—causing us to seem to see it. But however natural and human this temptation is, it must be resisted. We can be caused to seem to see something by something that shares no features with the illusory object. (Remember Ebenezer Scrooge saying to Marley's ghost: "You may be an undigested bit of beef, a blot of mustard, a crumb of cheese, a fragment of an underdone potato. There's more of gravy than of grave about you, whatever you are!") Many would insist that there has to be a ghost-shaped intermediary in the causal chain between blot of

¹ The material in the previous five paragraphs is adapted from Dennett (2013).

mustard and belief in Marley, but Scrooge might be right in addressing his remark to the cause of his current condition, and be leaving nothing Marley-shaped out.) And as for the idea that without being *rendered* such contents are causally impotent, it is simply mistaken, as a thought experiment will reveal. Suppose we have a drone aircraft hunting for targets to shoot at, and suppose that the drone is equipped with a safety device that is constantly on the lookout for red crosses on buildings or vehicles—we don’t want it shooting at ambulances or field hospitals! With its video eye it takes in and transduces (into digital bit streams) thirty frames a second (let’s suppose) and scans each frame for a red cross (among other things). Does it have to project the frame onto a screen, transducing bit streams into colored pixels? Of course not. It can make judgments based on un-transduced information—in fact, it can’t make judgments based on anything else. Similarly your brain can make judgments to the effect that there is a red stripe out there on the basis of spike train patterns in your cortex, and then act on that judgment (by causing the subject to declare “I seem to see a red stripe,” or by adjusting an internal inventory of things in the neighborhood, or ...). (I am deliberately using the word “judgment” for the drone’s discriminations and the brain’s discriminations; I have elsewhere called such items micro-takings or content-fixations. The main point of using “judgment” is to drive home the claim that these events are *not* anything like the exemplification of properties, intrinsic or otherwise. They are not qualia, in other words. Qualia—as typically conceived—would only get in the way. Don’t put a weighty LED pixel screen in a drone if you want it to detect red crosses, and don’t bother installing qualia in a brain if you want it to have color vision. Whatever they are, qualia are unnecessary and may be jettisoned without loss.)

So the familiar idea (familiar in the context of Block’s proposed distinction between access consciousness and phenomenal consciousness) that phenomenal consciousness (= qualia) is the basis for access consciousness (= judgments about qualia, qualia-guided decisions,

etc.) is backwards.² Once the discerning has happened in the untransduced world of spike trains, it can yield a sort of Humean projection—of a red stripe or red cross or just red, for instance—into “subjective space.”

But what is this subjective space in which the projection happens? Nothing. It is a theorist’s fiction. The phenomenon of “color phi” nicely illustrates the point. When shown, say, two disks displaced somewhat from each other, one sees the apparent motion of a single disk—the phi phenomenon that is the basis of animation (and motion pictures in general). If the disks are of different colors—the left disk red and the right disk green, for instance—one will see the red disk moving rightward and changing its color to green in mid-trajectory. How did the brain “know” to move the disk rightward and switch colors before having access to the green disk at its location? It couldn’t (supposing precognition to be ruled out). But it could have Bayesian expectations of continuous motion from place to place that provoke a (retrospective) expectation of the intermediate content, and this expectation encounters no disconfirmation (if the timing is right), which suffices to establish in reality the illusory sequence in the subject’s manifest image. So the visual system’s *access* to the information about the green disk is causally prior to the “*phenomenal*” motion and color change. Here is a diagram of color phi

2 I once had an occasion to point out this prospect to Block. He had just participated in a laterality test, to see how strongly lateralized for language his brain was. There are two oft-used ways of testing this: with dichotic headphones, which send different words into each ear, where the subject is asked to identify the word heard (typically you only hear one of them!). A second, visual test involves looking at a center target on a screen and having a word or non-word (e.g., “flum” or “janglet”) flashed briefly in either the left or right visual field. The subject presses the word button or the non-word button and latencies and errors are recorded. If you are strongly lateralized left (your left hemisphere is strongly dominant for language and does most of the work of language processing), you are faster and more accurate on words and nonwords flashed to the right hemifield. Ned had taken the visual test, and I asked him what he had learned. He was, he said, strongly lateralized left for language, like most people, and he added “the words flashed on the left actually seemed blurry!” I asked him whether the words seemed blurry because he noted the difficulty he was having with them, or whether he had the difficulty because the words were blurry. He acknowledged that he had no introspective way of distinguishing these two hypotheses. Supposing that Block doesn’t have some remarkable problem with his eyes, in which the left half of each lens is occluded or misshapen, producing a blur on the left side of his retinas, it is highly likely that the blurriness he seemed to experience was an effect of his felt difficulty in responding, not the cause of this difficulty.

from [Consciousness Explained](#) (and [Dennett & Kinsbourne 1992](#)):

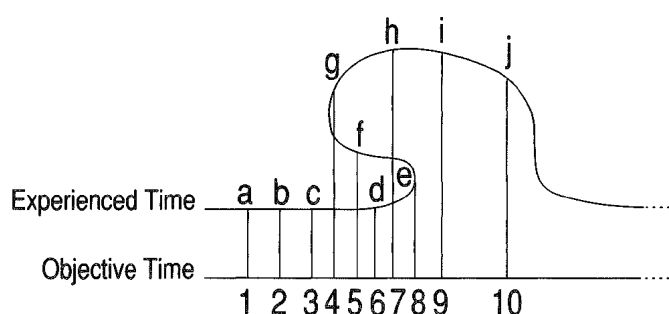


Figure 2: Superimposition of subjective and objective sequences.

In order to explain “temporal anomalies” of conscious experience, we need to appreciate that not only do we not have to represent red with something red, and round with something round; we don’t have to represent time with time. Recall my example “Tom arrived at the party after Bill did.” When you hear the sentence you learn of Tom’s arrival before you learn of Bill’s, but what you learn is that Bill arrived earlier. No revolution in physics or metaphysics is needed to account for this simple distinction between the temporal properties of a representation and the temporal properties represented thereby. It is quite possible (in color phi, for instance) for the brain to discern (in objective time) first one red circle (cat time 3) and then a green circle (fat time 5) displaced to the right, and then to (mis-)represent an intermediate red-turning-green circle (eat time 8) yielding the subjective judgment of apparent motion with temporally intermediate color change. Here our Bayesian probabilistic anticipator is caught in the act, jumping to the most likely conclusion in the absence of any evidence. Experienced or subjective time doesn’t line up with objective time, and it doesn’t have to. The important point to remember from the diagram is that the subjective time sequence is NOT like a bit of kinked film that then has to be run through a projector somewhere so that c is followed by e is followed by f in real time. It is just a theorist’s diagram of how subjective time can relate to objective time. Subjective time is not a further real component of the causal picture. No

rendering is necessary, the judgment is already in, and doesn’t have to be re-presented for another act of judging (in the Cartesian Theater).

The temptation to think otherwise may run deep, but it is fairly readily exposed. Consider fiction. Sherlock Holmes and Watson seem real when one is reading a Conan Doyle mystery—as real as Disraeli or Churchill in a biography. When Sherlock seems real, does this require him and his world to be rendered somewhere, in—let’s call it—*fictoplasm*? No. There is no need for a medium of fictoplasm to render fiction effective, and there is no need for a mysterious medium, material or immaterial, to render subjective experience effective. No doubt the temptation to posit the existence of fictoplasm derives from our human habit, when reading, of adding details in imagination that aren’t strictly in the book. Then, for instance, when we see a film of the novel, we can truly say “That’s not how I imagined Holmes when I read the book.”

Isn’t such rendering in imagination while reading a novel a case of *transduction* of content from one medium (written words as seen on the page) into another (imagined events as seen and heard in the mind’s eye and ear)? No, this is not transduction; it is, more properly, a variety of *translation*, *effortlessly expanding the content thanks to the built-in Bayesian prediction mechanisms*. We could construct, for instance, a digital device that takes problems in plane geometry presented in writing (“From Euclidean axioms prove the Pythagorean Theorem.”) and solves them through a process that involves making Euclidean constructions, with all the sides and angles properly represented and labeled, and utilizing them in the proof. The whole process from receipt of the problem to delivery of the called-for proof (complete with printed-out diagrams if you like), is conducted in a single medium of digital bit strings, with no transduction until the printer or screen is turned on to render the answer. (A more detailed description of this kind of transformative process without transduction is found in my discussion (1991) of how the robot Shakey discriminated boxes from pyramids.)

Consider [Figure 2](#) above. Does the access/phenomenal consciousness distinction get depicted therein? If so, access consciousness should be identified with the objective time line, and phenomenal consciousness (if it were something real in addition to access consciousness) would be depicted in the line that doubles back in time. The content feature that creates the kink is an effect of a judgment or discernment that came later in objective time than the discernment of the green circle at time 5. It is because the brain already had access to red circle, then green circle that it generated a representation (but not a *rendering*) of the in-between red-turning-green circle as an elaborative effect.³

5 Why do qualia seem so simple and ineffable?

Qualia seem atomic to introspection, unanalyzable simples—the smell of violets, the shade of blue, the sound of an oboe—but this is clearly an effect of something like the resolution of our discernment machinery.

If our vision were as poorly spatially resolved as our olfaction, when a bird flew by, the sky would suddenly “go all birdish,”—that peculiar, indescribable birdishness that one would experience in the visual presence of birds. And this resolution is variable: music lovers and wine enthusiasts and others can train up their ear and their palate and come to distinguish, introspectively, the combining elements of what used to seem atomic and unanalyzable. [David Huron \(2006\)](#), has done some ingenious work teasing out and explaining the combinations of neuroarchitectonic properties that explain the otherwise ineffable characteristic qualia of scale tones (the way *do* sounds different from *re* and *mi* and *so*). It turns out that these “qualia” are actually highly structured properties of neural representations. The explanation, needless to say, is ultimately in the medium of spike trains.

But why should the resolution (if that is the right term) be so low? Why should our brains ignore so much detail in the representations to which “we” have “access”? [Minsky](#)

³ Thanks to David Gottlieb for drawing my attention to this way of looking at access consciousness.

(1985), [Dennett \(1991\)](#), [Norretranders \(1999\)](#), [Metzinger \(2003\)](#), and others have said that it is the brain’s own access to its own complex internal activities that accounts for the simplicity. This is the brain’s effective user-illusion for itself, in much the way the desktop with its icons and various metaphors (click and drag, highlighted targets, etc.) is an elegantly designed user-illusion for laypeople who don’t need to know how their computers work.

The brain does not have a single internal witness or homunculus, but it does need something like a lingua franca to get the different and semi-independent subsystems to communicate with each other. (For instance, in the Global Neuronal Workplace model⁴ of [Dehaene et al. \(2006\)](#), and others, one should not take it for granted that the *local* meanings of spike train patterns—in the dorsal vision stream, say, or the olfactory bulb—are readily “understood” by all the elements to which some of these signals are broadcast.) I think there is bound to be some important truth in that theme, but it is only part of the story.

6 Whose access?

I think the more interesting suggestion is that the effective “we” when we talk about what “we” have access to, is, indeed, *we*—not just *I*, but *you and me*. It is, more particularly, *your* access to *my* mind that simplifies the information that *we* have access to!

The linguist [Stephen Levinson \(2006\)](#) has studied the remarkable language, Yéî Dnye, of the three thousand or so inhabitants of Rossel Island in the South Pacific—to the north of Papua New Guinea. It is a completely isolated language, unlike any other in the world in many regards. In particular, it is hideously complex, with:

the largest phoneme inventory (ninety distinct segments) in the Pacific, and many

⁴ Isn’t the Global Neuronal Workplace the derided Cartesian Theater after all? No, because what goes on there is not transduction-and-rendering, but informational integration: the coalition and consilience of competing elements. There is no transduction threshold that determines the time-of-entry “into consciousness”, and none of the multiple drafts competing in it are singled out as being conscious except retrospectively. This is the point of my admonition always to ask the Hard Question: “And then what happens?” ([Dennett 1991](#), p. 225)

sounds (such as doubly articulated labial coronal stops) that are either unique or rare in the languages of the world. Among the fifty-six consonants are many multiply articulated segments: e.g., /tɸɲm/ is a single segment made by simultaneously putting the tongue behind the alveolar ridge, trilling the lips, and snorting air through the nose. [...] Once the learner is past the sound hurdle, he or she faces another formidable obstacle. The language has an extremely complex system of verb inflection (with thousands of distinct inflectional forms). [...] In addition, substitute forms are used where the subject has been mentioned before, is close or visible, is in motion, or where the sentence is counterfactual or negative, thus providing well over a thousand possibilities [...]. (Levinson 2006, p. 20)

Levinson reports, not surprisingly, that “[h]ardly any mature individuals (such as non-native spouses) who have immigrated into the island community ever learn to speak the language, and children of expatriate Rossels do not fully acquire it from their parents alone.” His explanation is speculative, but plausible: a language, left to itself for centuries, will grow ever more complex, like an unpruned bush, simply because it can. The extreme isolation of Rossel Island over the centuries (for various geographic reasons) means that the language has hardly ever been confronted with non-native speakers of another language with whom communication is imperative, for one reason or another. The need for communication soon generates a small cadre of bi-lingual interpreters, and maybe also a pidgin (and maybe later a creole), and all of these alien interfaces work to simplify a language. The least learnable, most baroque (in the sense of exceeding the functional) features of the language are dropped under this pressure. We can see it happening with English today, with simplified dialects such as Emblish (as spoken at the European Molecular Biology Laboratory in Heidelberg) arising naturally and imperceptibly.

I would like to speculate that a similar process of gradual but incessant simplification has shaped the language we have available to explain and describe our minds to each other. Wittgenstein’s famous claim about the impossibility of a private language has not weathered the storms of controversy particularly well, but there are neighboring claims—empirical claims—that deserve consideration. Many years ago, [Nicholas Humphrey \(1987\)](#) made the point that has begun to attract adherents today:

While it is of no interest to a person to have the same kind of kidney as another person, it is of interest to him to have the same kind of mind: otherwise as a natural psychologist he’d be in trouble. Kidney transplants occur very rarely in nature, but something very much like mind-transplants occur all the time [...]. [So] we can assume that throughout a long history of evolution all sorts of different ways of describing the brain’s activity have been experimented with but only those most suited to doing psychology have been preserved. Thus the particular picture of our inner selves that human beings do in fact now have—the picture we know as ‘us’, and cannot imagine being of any different kind—is neither a necessary description nor is it any old description of the brain: it is the one that has proved most suited to our needs as social beings. That is why it works. Not only can we count on other people’s brains being very much like ours, we can count on the picture we each have of what it’s like to have a brain being tailor-made to explain the way that other people actually behave. Consciousness is a socio-biological product—in the best sense of socio and biological. (p. 18)

Chris Frith, for instance, has recently taken up the theme (in conversation) that consciousness has some features, because everything in consciousness has to be couched in terms that can be communicated to other people readily.

The ineffability barrier we all experience when trying to tell others what it is like to be

us on particular occasions is highly variable, not just between individuals, but over time within a single individual, as a result of formal or informal training. It plays a dynamic role in shaping the contents of our consciousness over time.⁵ (This would be true only for human consciousness, obviously.)

7 A thought experiment: Mr. Capgras

Finally, it might seem that whereas some subjective properties—cute, sweet, funny, sexy, the characteristic sounds of scale tones—might be accounted for in terms of Bayesian expectations about how one will be disposed to behave in their presence, the very simplicity of colors must block any attempt to treat them in a similar fashion. There is no way one expects to behave in the presence of navy blue, or pale yellow, or lime green. So it may seem, but this is itself an artifact of our penchant for thinking—as Hume famously did—of colors as simples. Hume was discountenanced by the notorious missing shade of blue, and found it ideologically inconvenient to suppose, as we now know, that color experience is in fact highly complex and compositional, and deeply anchored in dispositions of our perceptual systems.⁶ Moreover, color experiences are no more atomic than scale tone experiences, and give rise to all manner of expectations, which tend to go unnoticed, but can be thrown into sharp focus by a thought experiment: my fantasy about poor Mr. Clapgras, the man who wakes up to find all his emotional dispositions with regard to colors inverted while leaving intact his cognitive habits and powers (see [Dennett 2005](#), pp. 91–102, for a more detailed account, with objections considered and rebutted). Ex hypothesi, Mr. Clapgras identifies colors and sorts colors correctly (he does not suffer from the well-studied conditions color

anomia, or cerebral achromatopsia), but he finds the world disgusting, unbearable. Food looks just terrible to him now, and he has to eat blindfolded, since his emotional responses to all colors have shifted 180 degrees around the color circle ([Grush this collection](#)). He calls shocking pink “shocking pink” but marvels at the inappropriateness of its name. The only way we can explain his distress is by observing that he notices that something is wrong—which has to mean he was expecting something else. He is surprised that breaking a fresh egg into a frying pan on a sunny morning doesn’t bring a smile to his face, that a glimpse of his obnoxious neighbor’s lime green convertible doesn’t irritate him the way it used to do, that he feels no stirring of childhood patriotism when he sees the red white and blue waving in the breeze. Like the sufferers of Capgras delusion, poor Mr Clapgras senses a disturbance: something is very wrong, but it isn’t the evaporation of intrinsic internal properties.

8 Conclusion

The considerations I have raised in this essay are not new, but perhaps bringing them together as I have done will help show that a counter-intuitive theory like mine still has an advantage over some of the fantasies in which philosophers have recently indulged. It may well be, as [Paul Bloom \(2004\)](#) has suggested, that we are all “natural born dualists,” but just as eyeglasses can correct for myopia, natural-born or not, so science can correct for this innate cognitive disability. Intuitions to the contrary are important data, but should not be taken to indicate a limitation of science, as some have thought. In fact, if the best scientific theory of consciousness turns out not to be deeply counterintuitive at first, among the data it will have had to explain is why it took us so long to arrive at it.

Acknowledgements

Thanks to Michael Cohen, and to Andy Clark, and the rest of Dmitry Volkoff’s Greenland Consciousness and Free Will workshop (June 10-17, 2014), for editorial advice on this essay.

⁵ Note that I am not saying that our day-to-day consciousness wouldn’t occur in the absence of human company, but an implication of my speculation is that a Robinson Crusoe human, somehow raised from birth without human contact, would have subjectivity more inaccessible to us—once we discovered him and attempted to communicate with him—than the speech acts of the Rossel Islanders.

⁶ In [Cohen & Dennett](#), we point out that limbic or emotional responses to colors have to count as instances of “access” to color-representing states “however coarse-grained or incomplete, because such a reaction can obviously affect decision making or motivation” (2011, p. 5).

References

- Bloom, P. (2004). *Descartes' baby: How the science of child development explains what makes us human*. New York, NY: Basic Books.
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36 (3), 181-253. [10.1017/S0140525X12000477](https://doi.org/10.1017/S0140525X12000477)
- Cohen, M. A. & Dennett, D. C. (2011). Consciousness cannot be separated from function. *Trends in Cognitive Sciences*, 15 (8), 358-365. [10.1016/j.tics.2011.06.008](https://doi.org/10.1016/j.tics.2011.06.008)
- Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J. & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: A testable taxonomy. *Trends in Cognitive Sciences*, 10 (5), 204-211. [10.1016/j.tics.2006.03.007](https://doi.org/10.1016/j.tics.2006.03.007)
- Dennett, D. C. (1978). "What's the difference: Some riddles," (commentary on Puccetti and Dykes). *Behavioral and Brain Sciences*, 1 (3), 351-351.
- (1991). *Consciousness explained*. Boston, MA: Little, Brown and Company.
- (1994). Get real. *Philosophical Topics*, 22 (1-2), 505-568.
- (1995). "The path not taken," commentary on Ned Block, "On confusion about a function of consciousness". *Behavioral and Brain Sciences*, 18 (2), 252-253.
- (2005). *Sweet dreams: Philosophical obstacles to a science of consciousness*. Cambridge, MA: MIT Press.
- (2009). Darwin's 'strange inversion of reasoning'. *Proceedings of the National Academy of Sciences*, 106 (1), 10061-10065. [10.1073/pnas.0904433106](https://doi.org/10.1073/pnas.0904433106)
- (2013). Expecting ourselves to expect: The Bayesian brain as a projector (commentary on Clark, 2013). *Behavioral and Brain Sciences*, 36 (3), 209-210. [10.1017/S0140525X12002208](https://doi.org/10.1017/S0140525X12002208)
- Dennett, D. C. & Kinsbourne, M. (1992). Time and the observer: The where and when of consciousness in the brain. *Behavioral and Brain Sciences*, 15 (2), 183-247. [10.1017/S0140525X00068229](https://doi.org/10.1017/S0140525X00068229)
- Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin.
- Grush, R., Jaswal, L., Knoepfler, J. & Brovold, A. (2015). Visual Adaptation to a Remapped Spectrum: Lessons for Enactive Theories of Color Perception and Constancy, the Effect of Color on Aesthetic Judgments, and the Memory Color Effect. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Hume, D. (1739). *Treatise of human nature*. London, UK: John Noon.
- Humphrey, N. (1987). "The uses of consciousness" *The 57th James Arthur Lecture*. New York, NY: American Museum of Natural History.
- (2006). *Seeing red: A study in consciousness*. Cambridge, MA: Harvard University Press.
- (2011). *Soul dust: The magic of consciousness*. Princeton, NJ: Princeton University Press.
- Hurley, M., Dennett, D. C. & Adams, jr., R. B. (2011). *Inside jokes: Using humor to reverse-engineer the mind*. Cambridge, MA: MIT Press.
- Huron, D. (2006). *Sweet anticipation; Music and the psychology of expectation*. Cambridge, MA: MIT Press.
- Kail, P. J. E. (2007). *Projection and realism in Hume's philosophy*. Oxford, UK: Oxford University Press.
- Levinson, S. C. (2006). Introduction. In S. C. Levinson & P. Jaisson (Eds.) *Evolution and culture* (pp. 1-42). Cambridge, MA: MIT Press.
- Metzinger, T. (2003). *Being no one. The self-model theory of subjectivity*. Cambridge, MA: MIT Press.
- (2009). *The ego tunnel. The science of the mind and the myth of the self*. New York, NY: Basic Books.
- Minsky, M. (1985). *The society of minds*. New York, NY: Simon & Schuster.
- Norretranders, T. (1999). *The user illusion: Cutting consciousness down to size*. London, UK: Penguin Press Science.
- Prinz, J. (2012). *The conscious brain*. Oxford, UK: Oxford University Press.
- Sellars, W. (1962). Philosophy and the scientific image of man. In R. Colodny (Ed.) *Frontiers of science and philosophy* (pp. 35-78). Pittsburgh, PA: University of Pittsburgh Press.
- von Uexküll, J. (1957). A stroll through the worlds of animals and men: A picture book of invisible worlds. In C. H. Schiller (Ed.) *Instinctive behavior: The development of a modern concept*. New York, NY: International Universities Press.

Qualia explained away

A Commentary on Daniel C. Dennett

David H. Baßler

In his paper “Why and how does consciousness seem the way it seems?”, Daniel Dennett argues that philosophers and scientists should abandon Ned Block’s distinction between access consciousness and phenomenal consciousness. First he lays out why the assumption of phenomenal consciousness as a second medium is not a reasonable idea. In a second step he shows why beings like us must be convinced that there are qualia, that is, why we have the strong temptation to believe in their existence. This commentary is exclusively concerned with this second part of the target paper. In particular, I offer a more detailed picture, guided by five questions that are not addressed by Dennett. My proposal, however, still resides within the framework of Dennett’s philosophy in general. In particular I use the notion of intentional systems of different orders to fill in some details. I tell the counterfactual story of some first-order intentional systems evolving to become believers in qualia as building blocks of their world.

Keywords

Dispositions | Intentional systems | Predictive processing | Qualia | Zombic hunch

Commentator

David H. Baßler

davidhbassler@gmail.com

Johannes Gutenberg-Universität
Mainz, Germany

Target Author

Daniel C. Dennett

daniel.dennett@tufts.edu

Tufts University
Medford, MA, U.S.A.

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

The first of Rapoport’s Rules¹ for composing a critical commentary states that one should present the target view in the most charitable way possible (Dennett 2013a). Although I generally agree with many of Daniel Dennett’s

views, especially his argument against the existence of qualia (constituting the first part of the target paper), the diagnosis that there is the *zombic hunch*,² along with his strategy for explaining why it exists, the connection between qualia and predicted dispositions, was hard to grasp. Dennett presents the idea that when we talk about qualia, what we really refer to are our dispositions in earlier works (e.g., Dennett 1991). But the connection to predictive pro-

¹ Dennett named these rules after social psychologist and game theorist Anatol Rapoport. They are not to be confused with another “Rapoport’s Rule”, named after Eduardo H. Rapoport (cf. Stevens 1989). Here is the full list of Dennett’s Rapoport’s Rules:

1. “You should attempt to re-express your target’s position so clearly, vividly, and fairly that your target says, ‘Thanks, I wish I’d thought of putting it that way.’”
2. “You should list any points of agreement (especially if they are not matters of general or widespread agreement).”
3. “You should mention anything you have learned from your target.”
4. “Only then are you permitted to say so much as a word of rebuttal or criticism.”

(Dennett 2013a, p. 33)

² A philosophical zombie has nothing to do with any other sort of zombie. It behaves in *every* way like a normal person. The only difference is, that it lacks phenomenal experiences (though *ex hypothesi* it believes that it has phenomenal experiences). The zombic hunch is the intuition that a philosophical zombie would be different from us.

cessing is new (see also [Dennett 2013b](#)). There still seem to be some stepping stones missing, which I hope to fill in with my reconstruction. My goal is to provide a complete story that sticks as close to Dennett's argument as possible. This paper is not supposed to be a "rebuttal" or "criticism", but an "attempt to re-express [Dennett]'s position" (see footnote 1).

The structure of this commentary is as follows: in the [first](#) section I shall give a short outline of Dennett's explanation of why we have the zombic hunch. Since this involves the predictive processing framework, I shall give a very short introduction to this first. Following this, I present a short list of five questions that have not, in my opinion, yet been sufficiently addressed. In the [second](#) section I present an interpretation, or perhaps an extension, of Dennett's answers to these questions, by relying on the concept of an intentional system and using a strategy involving telling the counterfactual story of the evolution of some agents who end up believing in qualia (although *ex hypothesi* there are none). In the [third](#) section I shall analyze which features qualia should have, according to the beliefs of these agents, and show that there is at least a significant overlap with features many consider qualia to have.

I want to give a short justification for the unorthodox way of accounting for *beliefs* about *x* instead of for *x*'s existence itself. This is a general strategy found in other areas of Dennett's work. For example, he has asked, "Why should we think there is intentionality although there is none?" ([Dennett 1971](#)), "Why should we believe there is a god although there is none?" ([Dennett 2006](#)), and "Why should we think there is a problem with determinism and free will although there is none?" ([Dennett 1984, 2004](#)). Dennett's philosophy can in parts be seen as a therapeutic approach to "philosopher's syndrome"—"mistaking failures of imagination for insights into necessity" (e.g., [Dennett 1991](#), p. 401; [Dennett 1998a](#), p. 366)—by making it easier to see why we are convinced of the existence of something, even when there are good reasons to believe that it doesn't exist.

I want to draw attention to Hume's *Of Miracles* ([Hume 1995](#), X), where he states that the likelihood of a testimony about miracles being wrong is always greater than the likelihood of the miracle itself. This serves as a nice analogy for the case at hand: we might think of our own mind as a good "witness", but we already know too much about its shortcomings. So we should be suspicious when it cries out for a revolution in science or metaphysics, because this cry rests on the belief that something is missing, when no data but this very belief itself makes the demand necessary. Instead we should examine what else could have led our minds to form this conviction.

2 Dennett's proposal

In "Why and how does consciousness seem the way it seems?" Dennett gives an argument for why philosophers and scientists should abandon Ned Block's distinction between access consciousness and phenomenal consciousness, zombies, and qualia altogether. The argument is twofold: first Dennett lays down his argument for why the assumption of phenomenal consciousness as a second medium whose states are conscious experiences or qualia is "scientifically insupportable and deeply misleading" ([Dennett this collection](#), section 2). It is insupportable because there is simply no need to posit such entities to explain any of our behavior, so for reasons of parsimony they should not be a part of scientific theories (see also [Dennett 1991](#), p. 134). The assumption is deeply misleading because it makes us look for the wrong things, namely, the objects our judgments are about, rather than the causes of these judgments, which are nothing like these objects.

In a second step Dennett shows why creatures like us must be convinced that there are qualia, that is, why we have such a strong temptation to believe in their existence, *even though* there are no good reasons for this ([Dennett this collection](#), section 2 and 3; other places where Dennett acknowledges this conviction, the zombic hunch, are [Dennett 1999](#); [Dennett](#)

2005, Ch. 1; Dennett 2013a, p. 283). The following sections are exclusively concerned with this part of the target paper.

After completing the second step, Dennett explains why we ascribe qualia their characteristic properties—simplicity and ineffability (Dennett this collection, section 4 & 5). Although I also say something about this point (see section 4), Section 6 is an intuition pump (cf. e.g., Dennett 2013a) that will help the reader to apply Dennett’s alternative view to the experience of colors.

Before I present a short outline of Dennett’s second step, I want to briefly describe the predictive processing framework. This is necessary since both Dennett’s argument as well as my reconstruction make use of this framework. I shall not go into details of hierarchical predictive processing (PP) accounts here, since at least three papers in this collection (Clark, Hohwy, and Seth), as well as the associated commentaries (Madary, Harkness, and Wiese), are concerned with this topic and also offer ample references for introductory as well as further reading. I will instead give a very short description of the points that are most relevant to Dennett’s argument and recommend the above-mentioned papers and the references given there to the interested reader.

2.1 Predictive processing

In the PP framework, the brain refines an internal generative stochastic model of the world by continuously comparing sensory input (extero- as well as interoceptive) with predictions continuously created by the model. The overall model is spread across a hierarchy of layers, where the sensory layer is the lowest and each layer tries to predict (that is, to suppress) the activation pattern of the layer beneath it. The whole top-down activation pattern might be interpreted as a global hypothesis about the hidden causes of ongoing sensory stimulation. The difference between predicted and actual activation (*prediction error*) is what gets propagated up the hierarchy and leads to changes in the hypothesis. To be exact, this is only one possibility. Another is

that this leads to an action that changes the input in such a way that the prediction is vindicated (*active inference*, see e.g., Friston et al. 2011). However, although this aspect of PP—that it provides one formally-unified approach to perception and action—is a strength of the framework, it is not important here, given the context of this commentary. These changes are supposed to follow Bayes’ Theorem, which is why one might speak of Bayesian prediction (cf. e.g., Hohwy 2013).

The higher the layer in the hierarchy the more abstract the contents and the longer the time-scales or the predictive horizon. One example of a very abstract content is “only one object can exist in the same place at the same time” (Hohwy et al. 2008, p. 691, quoted after Clark 2013, p. 5).

One point to keep in mind is that, according to Hohwy (2014), this framework implies a clear-cut distinction between the mind and the world. That is, there is an *evidentiary boundary* between “where the prediction error minimization occurs” and “hidden causes [of the sensory stimulation pattern] on the other side” (Hohwy 2014, p. 7). I will come back to this point later in this commentary.

2.2 The outline of Dennett’s argument

1. Our own dispositions, expectations, etc. are part of the generative self-model instantiated by our brains. “We ought to have good Bayesian expectations about what we will do next, what we will think next, and what we will expect next” (Dennett this collection, p. 5)
2. When our brains do their job (described in (1)) correctly, i.e., there are no prediction-error signals, we misidentify dispositions of the organism with properties of another object. For instance, instead of attributing the disposition to cuddle a baby correctly to the organism having the disposition, our brain attributes “cuteness” to the baby.³ Color qualia

³ “Think of the cuteness of babies. It is not, of course, an ‘intrinsic’ property of babies, though it seems to be. [...] We expect to expect to feel the urge to cuddle it and so forth. When our expectations are fulfilled, the absence of prediction error signals is interpreted as confirmation that, indeed, the thing in the world with which we are in-

and other types of qualia also belong to this category.⁴

3. This means, under a personal level description, that we believe that there are properties *independent of the observer*, such as the cuteness of babies, the sweetness of apples, or the blueness of the sky, etc.
4. This is why it is so hard for us to doubt that qualia exist in the real world.

The crucial points seem to be (1) and (2). Before I lay out my interpretation I want to highlight some points that are not addressed in Dennett ([this collection](#)), but which are crucial if we are to have a complete picture. In the section *Our Bayesian brains*, I present a reconstruction that addresses these issues.

2.3 Five questions

1. **Why do we need to monitor our dispositions?** As noted in Dennett (2010), self-monitoring, in the sense of monitoring of our dispositions, values, etc., isn't needed unless one needs to communicate and to hide and share specific information about oneself at will. In his paper, Dennett does not address this issue, yet presupposes that "among the things in our *Umwelt* that matter to our well-being are ourselves". This is obvious if one reads "ourselves" as the motions of our bodies, but not so obvious if one includes things

interacting has the properties we expected it to have" (Dennett [this collection](#), p. 5).

- 4 The intuition pump of Mr. Clapgras in Dennett's section 6 is there to make the point that colors can be seen as dispositional properties of the organism rather than as properties of perceptual objects, in the same way as cuteness. Whether one is convinced by this or not, the intuitive problem seems to be the same: science tells us there are no properties like cuteness or color, while the zombic hunch tells us that this cannot be true. A more detailed discussion can be found in Dennett (1991, p. 375). I will not go into this here, but for the sake of argument I shall assume that this admittedly counter-intuitive categorization is acceptable. The reader's willingness to accept it might be helped by the following point given by Nicholas Humphrey, which reminds us that although at first thought colors do not *seem* to have action-provoking effects (like cuteness or funniness), after second thought one might think differently:
 "As I look around the room I'm working in, man-made colour shouts back at me from every surface: books, cushions, a rug on the floor, a coffee-cup, a box of staples—bright blues, reds, yellows, greens. There is as much colour here as in any tropical forest. Yet while almost every colour in the forest would be meaningful, here in my study almost nothing is. Colour anarchy has taken over."
 (Humphrey 1983, p. 149; quoted in Dennett 1991, p. 384).

like "what we will think next, and what we will expect next", as Dennett does (Dennett [this collection](#), p. 5). The next question is concerned with this latter form of self-monitoring:

2. **How is self-monitoring accomplished?** Hohwy (2014) refers to an evidential boundary in the predictive processing framework (see the section 2.1): there is a clear distinction between the mind/brain and the world (of which the body without the brain is a part), whose causal structure is yet to be revealed. Our expectations are part of our mind, which, if talk of the boundary is correct, does not have direct access to its own states *as* its own states—the mind is a black box to itself. So the prediction of its expectations needs to be indirect (just like the predictions of the causes of the sensory stimulation in general), and therefore the question arises how the self-monitoring of the mind is achieved according to Dennett. There is a further concern with self-monitoring, which one might call the "acquisition constraint" (cf. e.g., Metzinger 2003, p. 344):
3. **How did this self-monitoring evolve in a gradual fashion?** Large parts of [Breaking the Spell](#) are dedicated to making understandable how "belief in belief" could have evolved over the centuries, beginning long before the appearance of any religion. Dennett's goal here is quite similar: the explanation aims to make understandable how we came to believe in qualia, etc. But a step-by-step explanation is missing. I consider this form of the acquisition-constraint one of the most crucial for any satisfying explanation of this sort: each single step has to be understandable as one likely to have happened. One reason for this is that it would support a more fine-grained and mechanistic understanding; another is that it would satisfy the gradualism-constraint of Darwinism, which says that minds (just like anything else) "must have come into existence gradually, by steps that are barely discernible *even in retrospect*" (Dennett 1995, p. 200, emphasis in original).

Once we know why and how our brains accomplish the task of monitoring our dispositions and how they came to do so, one might still wonder why (as claimed in point 2, page 3) exactly these abstract properties of the organism would be misidentified as concrete properties of other things:

4. **Why do we misidentify our dispositions?** One of Dennett's central claims is that we misidentify our own dispositions, which leads to belief in qualia.⁵ Although misidentification seems to be ubiquitous (see superstition, religion, magic tricks, the rubber hand illusion—Botvinick & Cohen 1998; and even full body illusions—Blanke & Metzinger 2009) it nonetheless requires a special explanation in each case: is this a shortcoming of a system that has no disadvantages, or is it even something that benefits the system in some way (cf. McKay & Dennett 2009)? Keeping this last possibility in mind one might ask:

5. **Why are we so attached to the idea of qualia?** There seems to be something more that leads people to believe in qualia. There is the intuition that without qualia we would be very different—we would be “mere machines”, we could not *enjoy* things like a good meal or the smell of the air after it rains (a discussion of this characteristic of beliefs-about-qualia can be found in Dennett 1991, p. 383). Some might go further and say that our whole morality rests on the existence of qualia of pain and suffering (this worry is dealt with in Dennett 1991, p. 449). However, what I am concerned with here is not whether it is true that qualia are the basis of our morality, but why we should think them to be so. From the argument presented by Dennett it is not clear why we are so attached to the idea of qualia. It is not obvious why we do not react as disinterestedly to their denial as we did to the revelation that there is no

ether.⁶ But, as a matter of fact, we react differently: this is not like when any other entity, posited for theoretical reasons, is shown to not exist; it is as if without qualia we couldn't possibly be *us*.

3 An interpretation

3.1 Intentional Systems Theory

An important part of what follows is Intentional Systems Theory (IST). What is crucial here is that according to IST, all there is to being an agent in the sense of having beliefs and desires upon which to act is to be describable via a certain strategy: the *intentional stance*. The intentional stance is a “theory-neutral way of capturing the cognitive competences of different organisms (or other agents) without committing the investigator to overspecific hypotheses about the internal structures that underlie the competences” (Dennett 2009, p. 344). If one predicts the behavior of an object via the intentional stance, one presupposes that it is optimally designed to achieve certain goals. If there are divergences from the optimal path, one can, in a lot of cases, correct for this by introducing abstract entities or false beliefs. Since there are presumably no 100%-optimally-behaving creatures in the world, every intentional profile (a set of beliefs and desires), generated via adoption of the intentional stance, contains a subset of false beliefs.⁷ It seems that humans have a “generative capacity [to find the patterns revealed by taking the intentional stance] that is to some degree innate in normal people” (Dennett 2009, p. 342). I will come back to this point and its connection to PP in the next section.

Let us assume for the sake of argument that IST gives a correct explanation of what it is to be an agent (in the sense of someone who has beliefs and desires and acts according to

⁵ What qualia are [...] are just those complexes of dispositions. When you say ‘This is my quale,’ what you are singling out, or referring to, whether you realize it or not, is your idiosyncratic complex of dispositions. You seem to be referring to a private, ineffable something-or-other in your mind's eye, a private shadshade of homogeneous pink, but this is just how it seems to you, not how it is. (Dennett 1991, p. 389).

⁶ This property of the beliefs is acknowledged in Dennett (2005), p. 22, fn 18: “[The Zombic Hunch] is visceral in the sense of being almost entirely arational, insensitive to argument or the lack thereof”.

⁷ See Dennett (1987) for an elaborate discussion of the intentional stance and its implications, Dennett 1998b for the ontological status of beliefs and desires, Bechtel (1985) for another interesting interpretation, and Yu & Fuller (1986) for a discussion of the benefits of treating beliefs and desires as abstracta.

them), and that PP allows us to see how an agent can be implemented on the “algorithmic level”(see Dennett’s discussion in [Dennett 1987](#), p. 74, where he refers to the IST as a “competence model”). Whenever I say that an agent believes, wants, desires, etc. something I mean it in exactly the sense found in IST.

Intentional systems can be further categorized by looking at the content of their beliefs, e.g., a second-order intentional system is an intentional system that has beliefs and/or desires about beliefs and/or desires, that is, it is itself able to take an intentional stance towards objects ([Dennett 1987](#), p. 243). A first-order intentional system has (or can be described as having) beliefs and desires; a second-order intentional system can ascribe beliefs to others and itself. If something is a second-order intentional system it harbors beliefs such as “Peggy believes that there’s cheese in the fridge”. But taking the intentional stance towards an object is an ability that comes in *degrees*. I now want to describe what one might call an intentional system of 1.5^{th} order, an intermediate between first- and second-order intentional systems. This is a system that is not able to ascribe full-fledged desires and beliefs with arbitrary contents to others or itself. We, as intentional systems of high order, have no difficulty in ascribing beliefs and desires with very arbitrary contents, such as “She wants to ride a unicorn and believes that following Pegasus is a good way to achieve that goal”. But the content of beliefs and desires that such an intentional system of 1.5^{th} order can ascribe should be constrained in the following way:

1. An intentional system of 1.5^{th} order is able to ascribe desires only in a very particular and concrete manner, i.e., actions that the object in question wants to perform with certain particular existing objects, that the system itself knows about (e.g., the desire to eat the carrot over there), but not goals directed at nonexistent objects, described by sentences like “he wants to build a house”, or objects the ascriber itself does not know about.
2. It is only able to ascribe beliefs to others that it holds itself. That means it is able to

take the basic intentional stance with the default assumption that the target object in question believes whatever is true (if we assume the ascriber’s beliefs are in fact all true), but lacks the ability to correct the ascriptions if it leads to wrong predictions for the behavior of the target. A real-world example can be found in [Marticorena et al. \(2011\)](#): rhesus macaques in a false belief task can correctly predict what a person will do, given that the person knows where the object is hidden and they have seen the person getting to know this. They can also tell when a person doesn’t have the right knowledge, but they cannot use this information to make a prediction about where the person will look.

The implementation of such an intermediate between first- and second-order intentional systems can be easily imagined following predictive coding principles, as I will soon show. Following this, I argue that this sets down the basic fundamentals for systems evolving from this position to be believers in qualia, etc.

The reason for introducing this idea is that I want to show how, given predictive processing principles and a certain selection pressure, a 1.5^{th} -order-intentional-system might develop from a first-order-intentional-system. In a next step, I will argue that under an altered selection pressure such a system might become a full-fledged n^{th} -order-intentional-system, where n is greater or equal to two. Systems evolving in such a way, as I will describe, are bound to believe in the existence of something like qualia. In some sense this is only a just-so story, but the assumed selection pressures are very plausible, and the empirically-correct answer might not be too far away from this.

3.2 Our Bayesian brains⁸

To see how the pieces fit together imagine the situation of some first-order intentional systems, agents, which are the first of their kind. They act according to their beliefs and desires. They do so because the generative models im-

⁸ This section takes strong inspiration from Wilfrid Sellars’ section “Our Rylean Ancestors” in [Sellars \(1963, p. 178\)](#).

plemented in their brains generate a sufficient number of correct predictions about their environment for them to survive and procreate. They do a fairly good job of avoiding harms and finding food and mates. Since they are first-order intentional systems, the behavior of their conspecifics amounts to unexplained noise to them, because they are unable to predict the patterns of most of their behavior (which is what makes them *merely* first-order intentional systems), though they might well predict their behavior as physical objects, e.g., where someone will land if she falls off a cliff, for instance.

When resources are scarce, this leads to competition between these agents and it becomes an advantage to be able to predict the behavior of one's conspecifics. This behavior is by definition pretty complex (they are intentional systems), but one can get some mileage out of positing the following regularity: some objects in the world have properties that lead to predictable behavior in agents, e.g., if there is an apple tree this will lead to the agents approaching it, if they are sufficiently near, etc., whereas if there is a predator, they will run from it, etc. Their model of the world is populated by properties of items that allow the (arguably rough) predictions of *agent behavior*. One might indeed say that the desires of the agents are *projected*⁹ onto the world.¹⁰ Those who acquire this ability are now 1.5th order intentional systems (see above; monkeys and chimpanzees might turn out to be such, see

Roskies [this collection](#)).¹¹ However, findings in this area are controversial. See [Lurz 2010](#)), since they can predict the behavior of others, given that their behavior is indeed explainable via reference to actually-existing objects, such as apples or potential sexual partners. In addition to these properties, there is a new category of objects in “their world”: beings that react to these properties in certain ways.¹²

In a next step we might suppose that a system of communication or signaling evolves (the details are not important), turning our intentional systems of 1.5th order into communicative agents. As communicative beings they have an interest in hiding and revealing their beliefs according to the trustworthiness of others and their motives (cf. [Dennett 2010](#)). That is, any of those beings needs to have access to what it itself will do next, so that they can hide or share this information, depending on information about the other. One might think of hiding the information about one's desire to steal some food, and so on.

This is a situation where applying the predictive strategy that was formerly only used to explain the behavior of others to *oneself* becomes an advantage for each of the agents.¹³ Agents like this believe in the existence of a special kind of special kind of properties, i.e., they predict their *own* behavior on the basis of generative models that posit such properties: they believe that they approach apples *because* they are *sweet*, cuddle babies *because* they are *cute*, laugh about jokes *because* they are *funny*. Applying the strategy to their own behavior puts them in the same category (according to the generative model) as the others: they are unified objects that react to cer-

9 What I mean by “project” is that instead of positing an inner representation whose content is “I (the system in question) want to eat that apple” and whose function is a desire, along with correct beliefs about the current situation, what is posited is an eat-provocative property of the apple itself. Both theoretical strategies allow for the prediction of the same behavior. The crucial difference is that attributing new properties to objects that are already part of the model is a simpler way of extending the model than positing a complex system of internal states to each agent. Thus it is also more likely to happen. It's definitely much simpler than extending the model to incorporate all the entities that explain the behavior on a functional level (i.e., all the neurons, hormones etc.). It is successful to the same extent the intentional stance is successful, that is, in an arguably noisy way, but still successful enough to gain an advantage (since *ex hypothesi* all the conspecifics are intentional systems).

10 This is very close to Gibson's affordances (e.g., [Gibson 1986](#)) in that “values and meanings are external to the perceiver” (p. 127) and in a couple of other respects (*ibid.*). It is, however, different in that the postulated properties serve to predict the behavior of *others* and not to guide the behavior of the organism itself. For the relation between Gibsonian affordances and predictive processing see e.g., [Friston et al. \(2012\)](#).

11 “[R]ecent work on non-human primate theory of mind suggests that monkeys and chimpanzees have a theory of mind that represents goal states and distinguishes between knowledge and ignorance of other agents (the presence and absence of contentful mental representations), even if it fails to account for misrepresentation.” ([Roskies this collection](#), p. 12).

12 The selection of goals and other cognitive capabilities, etc., is all placed outside of the target object (see [footnote 9](#)). It will approach the object that has the highest attraction value, given that there is no object with a higher repulsion value, i.e., there is no internal selection process represented *as* internal selection. What makes other agents special objects, in this model, is that they react to properties that no other things react to, not that they have an internal life that is somehow special.

13 Notice that according to PP, there is no shortcut to be taken: the mind is a black box to itself—it has to infer its own properties just as any others.

tain properties, not a bunch of cells trying to live among one another.¹⁴

The agent-models of these beings might improve by integrating the fact that sometimes it is useful to posit non-existing entities or omit existing entities in order to predict the behavior of a given conspecific (think of subjects in the false belief-task looking in the wrong box). By this the concept of (false) beliefs arises. One can imagine how they further evolve into full-fledged second and higher-order intentional systems, in an arms-race for predicting their fellows.¹⁵

A further step: they develop sciences like we did and will come to have a scientific image of the world, which contains no special simple properties of objects that cause “agents” to behave in certain ways. They come to the conclusion that the brain does its job without taking notice of properties like cuteness or redness, “instead relying” on computations, which take place in the medium of spike trains and nothing but spike trains (cf. [target](#), section 1). Their everyday predictions of others and most importantly of themselves still rely on the posited properties. And some might wonder whether there isn’t something missing from the scientific image.

According to the scientific image, they, as biological organisms, react to photons, waves of air, etc., but these are not the contents of their own internal models employed in solving the continuous task of predicting themselves. The simplest things they react to seem to be colors and shapes, (perceived) sounds, etc. The reaction towards babies is explained via facial proportions and the like, but this is far from what their generative models “say”, which is “the reaction to babies is caused by their cuteness”.

They begin to build robots, which react to babies like they do. They say things like, “all this robot reacts to are the patterns in the baby’s face, the proportions one can measure;

but although it reacts like we do, it does not do so because of the baby’s cuteness”. Of course only non-philosophers might say that science misses a property of the baby, but philosophers still see that there is *something* missing, and since cuteness is not a property of the outside world, they conclude that it must be a property of the agents themselves.

This seems to me to be the current situation. We have the zombic hunch because it seems to us that there is something missing and it seems so because our generative models are built upon the assumption that there are properties of things out there in the world to which systems like us react in certain ways. We never consider others like us to be zombies because they are agents like us or better: we are systems like them. We dismiss robots because we know they can only react to measurable properties, which do not *seem* to us to be the direct cause of *our* behavior.

4 An analysis

Is it true that properties such as cuteness do not correspond to anything? In a sense it is false to deny that any such correspondence exists: such properties do correspond to the cuddle-provocativeness of a baby, the eating-provocativeness of an apple, etc., *as a cause of the behavior of agents*. They are “lovely” properties ([Dennett 1991](#), p. 379), and there is a way to measure them: we can use ourselves as detectors. But the reason we, intuitively, do not accept a robot as a subject like ourselves is because we know how the robot does it: we know that it calculates, maybe even in a PP-manner—we know that it does not react directly to the properties that seem to exist and that seem to count. Neither do we, or the beings described above. But their own prediction of themselves treats such complex properties as simple, because there is nothing to be gained by being more precise than is necessary for *sufficiently* accurate prediction.¹⁶

This is my reconstruction of Dennett’s claim that the mind projects its dispositions

¹⁴ This is where one might speak of the origin of a self-model ([Metzinger 2003](#)) in some sense, where there is not only a model of the body (built up by proprioceptive inputs) but also a model of the self as having (primitive) goals, at least in any given moment.

¹⁵ Maybe language plays an important part in this further development as an external scaffold (cf. [Clark 1996](#); [Dennett 1994](#)). One fact supporting this view is that monkeys do not seem to be able to understand the concept of false belief (and therefore the concept of belief) (cf. [Martcorena et al. 2011](#), but also [Lurz 2010](#) for an overview of this debate).

¹⁶ This is also true of affordances (see e.g., [Gibson 1986](#), p. 141).

onto the world via Bayesian prediction. I want to draw attention to some of the features ascribed to those properties that this story predicts:

1. These properties are “given directly” to a person

The overall generative model depicts the whole organism as a unified object that reacts *directly* to the posited properties in the world. Any system that represents itself in such a way is bound to believe that there are properties of the world given directly to the object, which it takes to be itself. In subpersonal terms this object and these properties, as well as their relation to each other, are postulated entities that explain the sensory input. For instance, the fact that others talk about the system as someone with beliefs and desires (which is rooted in the same principle) can be explained by predicting itself in the same way.

2. These properties are irreducible to physical, mechanical phenomena.

Since the generative model does not depict these properties as built up from simpler ones, but simply posits them to predict lower-level patterns, these properties don't seem (to the system) to be reducible to other properties.

3. These properties are atomic, i.e., unstructured.

There are as many posited properties as there are distinct dispositions to be tracked. This also explains why one can learn to find structure in formerly unstructured qualia (cf. [Dennett 1991](#), p. 49) once new discriminative behavior is learned.

4. These properties are important to our lives/beings as humans/persons

This felt importance is obvious, given the putative role they play in the explanation provided by the generative model. These properties seem to be the causes of all our behavior: if one did not feel the painfulness of a pain, one would not scream; if one did not sense the funniness of a joke, one would not laugh, etc. Since the model is still needed for interacting with others, despite theoret-

ical advances in the sciences this felt importance of qualia to our lives is very difficult to overcome.

5. These properties are known to every living human being; it is not possible to sincerely deny their existence

This is due to the fact that our brains predict the behavior of others via a model that posits direct interaction between “agents” and first-order, non-relational object properties—the entities that are then named “qualia”.

This list has considerable overlap with lists of features ascribed to qualia (e.g., [Metzinger 2003](#), p. 68; [Tye 2013](#)), lending support to the thesis that we don't need a revolution in science to accommodate qualia, but rather a change in perspective: we might look at the creatures described above and see that “[t]hey are us” ([Dennett 2000](#), p. 353).

5 Conclusion

I have given an interpretation of Dennett's theory of why there seems to be something more to consciousness than science can explain. My aim was to thereby address crucial questions, while sticking as closely to Dennett's philosophy as possible. The answer is a just-so story that shows how (plausible) selection pressures lead to beings that cannot help but believe that they are *more* than just “moist robots” ([Dennett 2013a](#), p. 49)—because some important entities seem to be missing from the scientific description.

This story answers the questions why and how beings like us monitor their dispositions, and how this ability could have evolved. It also offers an answer as to why we don't recognize them as representations of our dispositions and why qualia are unlike other theoretical entities in that they are important for what we consider ourselves to be. The notion of an intermediate between first- and second-order intentional systems was introduced as a new conceptual instrument for satisfying the acquisition constraint and to lay the fundamentals for the belief in mind-independent simple properties that dir-

ectly cause the behavior of agents. This in turn is the basis for the belief in qualia as intrinsic properties of experience.

This story might not provide an “insight into necessity” (cf. Dennett 1991, p. 401), but I am happy if it contributes to showing and clarifying a possibility: although it may *seem* that our best hypothesis for accounting for our belief in qualia is that they actually exist, this hypothesis might still be explained away.

Acknowledgements

I want to thank Thomas Metzinger and Jennifer Windt for the unique opportunity to participate in this project. I am also very grateful for the helpful remarks they and two anonymous reviewers gave to an earlier version of this paper.

References

- Bechtel, W. (1985). Realism, instrumentalism, and the Intentional Stance. *Cognitive Science*, 9 (4), 473-497. [10.1207/s15516709cog0904_5](https://doi.org/10.1207/s15516709cog0904_5)
- Blanke, O. & Metzinger, T. (2009). Full-body illusions and minimal phenomenal selfhood. *Trends in Cognitive Sciences*, 13 (1), 7-13. [10.1016/j.tics.2008.10.003](https://doi.org/10.1016/j.tics.2008.10.003)
- Botvinick, M. & Cohen, J. (1998). Rubber hands “feel” touch that eyes see. *Nature*, 391 (756). [10.1038/35784](https://doi.org/10.1038/35784)
- Clark, A. (1996). Linguistic anchors in the sea of thoughts. *Pragmatics & Cognition*, 4 (1), 93-103. [10.1075/pc.4.1.09cla](https://doi.org/10.1075/pc.4.1.09cla)
- (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36 (3), 181-204. [10.1017/S0140525X12000477](https://doi.org/10.1017/S0140525X12000477)
- (2015). Embodied prediction. In T. Metzinger & J. W. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Dennett, D. C. (1971). Intentional systems. *Journal of Philosophy*, 68, 87-106.
- (1984). *Elbow room. The varieties of free will worth wanting*. Oxford, UK: Clarendon Press.
- (1987). *The Intentional Stance*. Cambridge, MA: MIT Press.
- (1991). *Consciousness explained*. New York, NY: Back Bay Books/Little, Brown and Company.
- (1994). The Role of Language in Intelligence. In J. Khalfa (Ed.) *What is Intelligence? The Darwin College Lectures*. Cambridge, UK: Cambridge University Press. [10.1075/pc.4.1.09cla](https://doi.org/10.1075/pc.4.1.09cla)
- (1995). *Darwin’s dangerous idea: Evolution and the meanings of life*. New York, NY: Simon Schuster Paperbacks.
- (1998a). Self-portrait. *Brainchildren: Essays on designing minds* (pp. 355-366). Cambridge, MA: MIT Press.
- (1998b). Real patterns. *Brainchildren: Essays on designing minds* (pp. 95-120). Cambridge, MA: MIT Press.
- (1999). The zombic hunch: Extinction of an intuition. *Royal Institute of Philosophy Millennial Lecture*
- (2000). With a little help from my friends. In D. Ross, A. Brooks & D. Thompson (Eds.) *Dennett’s Philosophy: A Comprehensive Assessment* (pp. 327-388). Cambridge, MA: MIT Press.
- (2004). *Freedom Evolves*. London, UK: Penguin Books.

- (2005). *Sweet dreams: Philosophical obstacles to a science of consciousness*. Cambridge, MA: MIT Press.
- (2006). *Breaking the spell. Religion as a natural phenomenon*. New York, NY: Penguin.
- (2009). Intentional Systems Theory. In B. P. McLaughlin, A. Beckermann & S. Walter (Eds.) *The Oxford handbook of philosophy of mind* (pp. 339-349). Oxford, UK: Oxford Handbooks Online.
- (2010). The evolution of why. In B. Weiss & J. Wanderer (Eds.) *Reading Brandom: On making it explicit* (pp. 48-62). New York, NY: Routledge.
- (2013a). *Intuition pumps and other tools for thinking*. New York, NY: W. W. Norton & Co..
- (2013b). Expecting ourselves to expect: The Bayesian brain as a projector. *Behavioral and Brain Sciences*, 36 (3), 29-30. [10.1017/S0140525X12002208](https://doi.org/10.1017/S0140525X12002208)
- (2015). Why and how does consciousness seem the way it seems? In T. Metzinger & J. W. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Friston, K., Mattout, J. & Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics*, 104 (1-2), 137-160. [10.1007/s00422-011-0424-z](https://doi.org/10.1007/s00422-011-0424-z)
- Friston, K. J., Shiner, T., FitzGerald, T., Galea, J. M., Adams, R., Brown, H., Dolan, R. J., Moran, R., Stephan, K. E. & Bestmann, S. (2012). Dopamine, affordance and active inference. *PLoS Computational Biology*, 8 (1), e1002327-e1002327. [10.1371/journal.pcbi.1002327](https://doi.org/10.1371/journal.pcbi.1002327)
- Gibson, J. J. (1986). *The ecological approach to visual perception*. Hillsdale, NJ: Erlbaum.
- Harkness, D. (2015). From explanatory ambition to explanatory power-A commentary on Jakob Hohwy. In T. Metzinger & J. W. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Hohwy, J. (2013). *The predictive mind*. Oxford, UK: Oxford University Press.
- (2014). The self-evidencing brain. *Noûs*, online. [10.1111/nous.12062](https://doi.org/10.1111/nous.12062)
- (2015). The neural organ explains the mind. In T. Metzinger & J. W. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Hohwy, J., Ropstorff, A. & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, 108 (3), 687-701. [10.1016/j.cognition.2008.05.010](https://doi.org/10.1016/j.cognition.2008.05.010)
- Hume, D. (1995). *An inquiry concerning human understanding*. London, UK: Pearson.
- Humphrey, N. (1983). *Consciousness regained*. Oxford, UK: Oxford University Press.
- Lurz, R. W. (2010). Belief attribution in animals: On how to move forward conceptually and empirically. *Review of Philosophy and Psychology*, 2 (1), 19-59. [10.1007/s13164-010-0042-z](https://doi.org/10.1007/s13164-010-0042-z)
- Madary, M. (2015). Extending the explanandum for predictive processing-A commentary on Andy Clark. In T. Metzinger & J. W. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Martcorena, D. C.W., Ruiz, A. M., Mukerji, C., Goddu, A. & Santos, L. R. (2011). Monkeys represent others' knowledge but not their beliefs. *Developmental Science*, 14 (6), 1467-7687. [10.1111/j.1467-7687.2011.01085.x](https://doi.org/10.1111/j.1467-7687.2011.01085.x)
- McKay, R. T. & Dennett, D. C. (2009). The evolution of misbelief. *Behavioral and Brain Sciences*, 32, 493-561. [10.1017/S0140525X09990975](https://doi.org/10.1017/S0140525X09990975)
- Metzinger, T. (2003). *Being no one. The self-model theory of subjectivity*. Cambridge, MA: MIT Press.
- Roskies, A. (2015). Davidson on believers: Can non-linguistic creatures have propositional attitudes? In T. Metzinger & J. W. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Sellars, W. (1963). *Science, perception and reality*. London, UK: Routledge & Kegan Paul Ltd..
- Seth, A. (2015). The cybernetic bayesian brain: From interoceptive inference to sensorimotor contingencies. In T. Metzinger & J. W. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Stevens, G. C. (1989). The latitudinal gradients in geographical range: How so many species co-exist in the tropics. *American Naturalist*, 133 (2), 240-256.
- Tye, M. (2013). Qualia. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy (Fall 2013 Edition)*
- Wiese, W. (2015). Perceptual presence in the Kuhnian-Popperian Bayesian brain—A commentary on Anil Seth. In T. Metzinger & J. W. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Yu, P. & Fuller, G. (1986). A critique of Dennett. *Synthese*, 66 (3), 453-476. [10.1007/BF00414062](https://doi.org/10.1007/BF00414062)

How our Belief in Qualia Evolved, and Why We Care so much

A Reply to David H. Baßler

Daniel C. Dennett

David Baßler's commentary identifies five unasked questions in my work, and provides excellent answers to them. His explanation of the gradual evolution of higher-order intentionality via a Bayesian account leads to an explanation of the persistence of our deluded belief in qualia.

Keywords

Belief in belief | Dispositions | Intentional systems | Qualia

Author

[Daniel C. Dennett](#)

daniel.dennett@tufts.edu

Tufts University

Medford, MA, U.S.A.

Commentator

[David H. Baßler](#)

davidhbassler@gmail.com

Johannes Gutenberg-Universität

Mainz, Germany

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität

Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University

Melbourne, Australia

David Baßler's commentary is a model of constructive criticism, not only pointing to weaknesses but offering persuasive repairs. I have just two points of minor correction to offer before turning to my understanding of his interesting proposals for extensions to my view, which I am inclined to adopt.

First, then, the quibbles. I am happy to see him endorsing my frequent tactic of asking not how to explain *x* but rather asking how to explain why we believe in *x* in the first place, but I think that this is a procrustean bed on which to stretch my concept of intentional sys-

tems. In [Dennett \(1971\)](#) I was indeed offering an account of intentionality that was *demoting*, in that intentionality was not seen as a feature that sundered the universe into the mental and physical (as Brentano and others had claimed), but I don't like to think of it as dismissing intentionality as a real phenomenon—though of course many have interpreted me that way. [Dennett \(1991\)](#) tried to correct that misconstrual, showing that the phenomena of intentionality are real in their own way—any beings that don't discover these patterns are missing something important in the world. That aside, I

love the use he makes of Hume on miracles to introduce his treatment of our minds as witnesses, just not very good witnesses; their testimony can be explained in ways that do not grant the truth of some of their most cherished claims. As he puts it, the assumption of phenomenal consciousness “is deeply misleading because it makes us look for the wrong things, namely, the objects our judgments are about, rather than the causes of these judgments, which are nothing like these objects” (Baßler [this collection](#), p. 2).

My other quibble is a similar elision I want to resist. He says: “Large parts of *Breaking the Spell* are dedicated to making understandable how ‘belief in belief’ could have evolved over the centuries, beginning long before the appearance of any religion” (Baßler [this collection](#), p. 4). This misidentifies higher order belief, beliefs about beliefs, with belief in belief. The former did indeed evolve gradually over the eons, and I find Baßler’s “just so story” about this gradual process enticing indeed, and will have more to say about it below, but belief in belief is a much younger (and almost always pernicious) phenomenon, which involves the deeply confused judgment that it is morally obligatory to try to get yourself to believe traditional nonsense when you know better. “If you don’t believe in God, you are immoral. Therefore you must strive to believe in God. Belief in God is a good thing to inculcate in our children and in ourselves.” Belief in belief didn’t arrive on the human scene until the proto-religions (which originally had no need for the concept) hit upon this obligation as a way of protecting their hegemony against the lures of competing dogmas. Some proto-religions were blithely ecumenical, adopting the gods and demons of their neighbors’ creeds as just another bit of lore about the big wide world, but this credulity could not long stand in the face of market competition and growing common knowledge about the objective world. Since many—probably most—people in the world now see through at least most of the nonsense, their persistent belief in belief is now a deplorable anachronism, a systematic source of hypocrisy. (A delightful cartoon in a recent *New Yorker* perfectly cap-

tures this folly. Two armies confront each other, flying identical banners; one mounted warrior says “There can be no peace until they renounce their Rabbit God and accept our Duck God.”)

As I say, these are quibbles I have to get off my chest. Now to Baßler’s substantive proposals. He organizes his commentary around five questions he says I haven’t properly asked, and he has answers to all of them. He’s right that these are gaps in my account. (1) Why do we need to monitor our dispositions? (2) How is self-monitoring accomplished? (3) How did this self-monitoring evolve in a *gradual* fashion? (4) Why do we misidentify our dispositions? (5) Why are we so attached to the idea of qualia?

His answers are constructed by taking on, for the sake of argument, my Intentional Systems Theory, and he gets it right, in all regards. Intentional Systems Theory (IST) presupposes, tactically, that any entity treated as an intentional system “is optimally designed to achieve certain goals. If there are divergences from the optimal path, one can, in a lot of cases, correct for this by introducing abstract entities or false beliefs.” IST is, as I say, a competence model that leaves implementation or performance questions unaddressed.¹

Then comes Baßler’s major novelty: the idea of an intermediate competence between mere first-order intentional systems—which have no beliefs about beliefs (their own or others’)—and full-fledged second-or-higher-order intentional systems—which can iterate the belief context. Such entities he calls (what else?) 1.5th order intentional systems (shades of [David Marr’s 1982](#) two-and-a-half-D sketch!). This is proposed to answer his first and second questions with a plausible and in principle testable evolutionary hypothesis. A system with only 1.5th order intentionality “is able to ascribe desires only in a very particular and concrete

¹ In this regard it is strikingly similar to the free energy principle as presented by [Hohwy \(this collection\)](#); both use the assumption of biofunctional optimizing as an interpretive lever to make sense of the myriad complexities of the brain, assigning to the brain a fundamental task of acquiring accurate anticipations of the relevant causes in the organism’s world. I have not yet been able to assess the costs and benefits of these two different ways of thinking of brains as future-producer: both are abstract, both court triviality if misused. This is a good topic for future work.

manner, i.e., actions that the object in question wants to perform with certain particular existing objects, that the system itself [the ascriber] knows about” (Baßler [this collection](#), p. 6). He is wise to choose basic desires (for food, mating opportunities, safety, . . .) as the intentional states ascribed in this precursor mentality, since they are so readily “observable” in the immediate behavior of the object, giving our pioneer mind-reader a quick confirmation that it’s on the right track, a small, gradual step for a Bayesian brain.

Now what selection pressures would favor such systems evolving gradually from mere first-order systems? To the primitive first-order systems, “the behavior of their conspecifics is unexplained noise to them.” But then they make some simple discoveries. When they see an apple tree, they approach it, *and so do their conspecifics*. If they see a predator, they run, as do their kin. “One might indeed say that the desires of the agents are projected onto the world”, Baßler says. Then, in a very substantive footnote that I wish were in the text—his footnotes contain much of value, and should not be passed over!—he adds: “What I mean by ‘project’ is that instead of positing an inner representation . . . whose function is a desire, along with correct beliefs about the current situation, what is posited is an eat-provocative property of the apple itself. Both theoretical strategies allow for the prediction of the same behavior. The crucial difference is that attributing new properties to objects that are already part of the model is a simpler way of extending the model than positing a complex system of internal states to each agent” (Baßler [this collection](#), p. 7, footnote 9). This answers question (3).

He then imagines, plausibly, that these 1.5th-order systems will evolve a system of communication, but this (as I and others have argued) necessarily involves hiding information from others, which involves having an internal cache of self-monitored knowledge one can choose to divulge or not, depending on circumstances. And this in turn—Baßler’s next major innovation—leads them to become “Agents [who] believe in the existence of a special kind of properties: they believe that they approach

apples because they are sweet, cuddle babies because they are cute, laugh about jokes because they are funny.” This primitive concept of causation serves them well, of course, and is just the sort of simplification to expect in a Bayesian brain, answering question (4).

Now for the icing on the cake, Baßler’s answer to question (5) about why we care about qualia. As he notes, “It is not obvious why we do not react as disinterestedly to their denial as we did to the revelation that there is no ether” (Baßler [this collection](#), p. 5). Here is his explanation: science comes along and starts to dismantle the handy manifest image, with all its Gibsonian affordances, and for those creatures capable of understanding science, a new problem arises: something is being taken away from them! All those delectable properties (and the abhorrent properties as well, of course). Philosophers “still see that there is something missing, and since cuteness is not a property of the outside world, they conclude that it must be a property of the agents themselves” (Baßler [this collection](#), p. 8). “We have the zombic hunch because it seems to us that there is something missing and it seems so because our generative models are built on the assumption that there are properties of things out there in the world to which systems like us react in certain ways. . . . We dismiss robots because we know they can only react to measurable properties, which do not seem to us to be the direct cause of our behavior” (*ibid.*).

This rings true to me, and I hadn’t seen this way of accounting for the persistence of the zombic hunch. Baßler proposes that “the reason we, intuitively, do not accept a robot as a subject like ourselves is because we know how the robot does it; we know that it calculates, maybe even in a PP manner—we know that it does not react directly to the properties that seem to exist and that seem to count” ([this collection](#), p. 9). He goes on to list five further features his account provides for. The properties we delusionally persist in “projecting” as qualia are (1) “‘given directly’ to a person”, (2) “irreducible to physical, mechanical phenomena”, (3) “atomic, unstructured”, (4) “important to our lives/beings as humans/persons”, and (5)

“known to every living human being; it is not possible to sincerely deny their existence” (Baßler [this collection](#), p. 9). I particularly like the way that his account explains why (4) is a feature: “These properties seem to be the causes of all our behavior: if one did not feel the painfulness of a pain, one would not scream; if one did not sense the funniness of a joke, one would not laugh, etc. Since the model is still needed for interacting with others, despite theoretical advances in the sciences this felt importance of qualia to our lives is very difficult to overcome” (Baßler [this collection](#), p. 9).

I see that my response consists in large measure of approving quotations from Baßler’s commentary! But that is as it must be; I want to confirm in detail and acknowledge the nice way his proposals dovetail with my account, expanding it into new territory, and helping me see what I have so far only dimly appreciated: just how valuable the new Bayesian insights are.

But let me end with a friendly amendment of my own. Baßler’s interpretation of my view is at one point a simplification, probably just for gracefulness of exposition, and perhaps meant itself as a friendly amendment, but I want to issue a caveat. Baßler takes me to be saying that, for such properties as cuteness and color, “we misidentify dispositions of the organism with properties of another object” ([this collection](#), p. 3) and goes on to have me holding that “This means, under a personal level description, that we believe that there are properties independent of the observer, such as the cuteness of babies, the sweetness of apples, or the blueness of the sky” (*ibid.*, p. 4). I want to put this slightly differently. It is not that there is nothing objective about babies that makes them cute (or of the sky that makes it blue) but just that these objective, observer-independent properties are themselves curiously dispositional: they are, as he notes at one point, what I have called “lovely” properties. They can only be *defined* relative to a target species of observers, such as normally sighted—not “color-blind”—human beings, as contrasted with tetrachromats such as pigeons, for instance. But their existence as properties is trivially objective and observer-independent. Thus rubies were *red* before color

vision evolved on this planet in the sense that if a time machine could take normal human beings back to the early earth, they would find rubies to be red. And some strata exposed by primordial earthquake faults would have been *visible*, to some kinds of eyes and not to others. Probably dinosaur babies were cute, since, as John Horner (1998) has argued, evidence strongly suggests that they were altricial, requiring considerable parental attention, and having the foreshortened skull and facial structure of prototypically cute juvenile animals, including birds. The science-endorsed properties, both external and internal, are so hugely different from what the manifest image makes them out to be, that it is a pickwickian stretch to say that science has discovered “what cuteness is” or “what color is,” but it is also deeply misleading to say that science has discovered that nothing is cute, or colored, after all. And so in a similar vein, I have to contend with how to occupy the awkward middle ground between denying that there are qualia at all, or saying that qualia are something real, but something utterly unlike what most people *think* (and philosophers *say*) qualia are.

1 Conclusion

Baßler has provided me with a plausible and testable extension of my Intentional System Theory with his innovation of a 1.5th-order intentional system, showing in outline how higher-order intentional systems might evolve from their more primitive ancestors. And he has also shown new ways of explaining a point that many people just cannot get their heads around. As my former student Ivan Fox (1989) once put it, “Thrown into a causal gap, a quale will simply fall through it.” See also Fox’s essay, “[Our Knowledge of the Internal World](#)” (1994) and [my commentary](#) on it (1994), which I discovered, on rereading just now, to be groping towards some of the points in Baßler’s commentary. I challenged Ivan Fox to “push further into the engineering and not just revel in the specs” (Dennett 1994, p. 510), and Baßler has done just that.

References

- Baßler, D. H. (2015). Qualia explained away: A commentary on Daniel C. Dennett. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Dennett, D. C. (1971). Intentional systems. *The Journal of Philosophy*, 68 (4), 87-106. [10.2307/2025382](https://doi.org/10.2307/2025382)
- (1991). Real patterns. *The Journal of Philosophy*, 88 (1), 27-51. [10.2307/2027085](https://doi.org/10.2307/2027085)
- (1994). Get real. *Philosophical Topics*, 22 (1 & 2), 59-106.
- Fox, I. (1994). Our knowledge of the internal world. *Philosophical Topics*, 22 (1 & 2), 59-106.
- Hohwy, J. (2015). The neural organ explains the mind. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Horner, J. R., Gorman, J., Henderson, D. & Blumer, T. L. (1998). *Maia: A dinosaur grows up*. Bozeman, MT: Museum of the Rockies, Montana State University.
- Marr, D. (1982). *Vision*. New York, NY: Freeman.

The Heterogeneity of Experiential Imagination

Jérôme Dokic & Margherita Arcangeli

Imagination is very often associated with the experienceable. Imagination is said to “re-create” conscious experiences. For instance, philosophers often talk of vision-like or audition-like imagination. How many varieties of experiential imagination are there, and how are they related? In this paper, we offer a detailed taxonomy of imaginative phenomena, based on both conceptual analysis and phenomenology, which contributes to answering these questions. First, we shall spell out the notion of experiential imagination as the imaginative capacity to re-create experiential perspectives. Second, we suggest that the domain of experiential imagination divides into objective and subjective imagination. In our interpretation, objective imagination comprises both sensory and cognitive imagination. In contrast, subjective imagination re-creates non-imaginative internal experiences of one’s own mind, including proprioception, agentive experience, feeling pain, and perhaps internal ways of gaining information about other types of mental states, such as sensory experience and belief. We show how our interpretation of the notion of subjective imagination differs from Zeno Vendler’s, who relies on an orthogonal distinction between two ways in which the self is involved in our imaginings. Finally, we show the relevance of our taxonomy for several important philosophical and scientific applications of the notion of imagination, including modal epistemology, cognitive resonance, mindreading and imaginative identification.

Keywords

Cognitive imagination | Cognitive resonance | Experiential imagination | External experience | Imagination | Imagination from the inside | Imaginative identification | Internal experience | Introspection | Mindreading | Motor imagery | Objective imagination | Recreative imagination | Self-involvement | Sensory imagination | Subjective imagination

1 Introduction

Many theorists have pointed out that imagination, or at least a salient type of imagination, is bound to the “experienceable”.¹ In this sense, we can imagine only what can be experienced. For

¹ This is a widespread claim that dates back to Plato and Aristotle and pervades the history of philosophy. See [White \(1990\)](#) for a good survey and a critical view of the standard picture (suggested by etymology) according to which imagination is akin to perception only. Among contemporary philosophers see also, for instance, [Wollheim \(1984\)](#), [Williams \(1976\)](#), [Casey \(1976\)](#), [O’Shaughnessy \(1980\)](#), [Vendler \(1984\)](#), [Peacocke \(1985\)](#), [Walton \(1990\)](#), [Mulligan \(1999\)](#), [Kind \(2001\)](#), [Currie & Ravenscroft \(2002\)](#), [Martin \(2002\)](#), [Noordhof \(2002\)](#), [Chalmers \(2002\)](#), [Carruthers \(2002\)](#), [McGinn \(2004\)](#), [Goldman \(2006\)](#), [Byrne \(2010\)](#).

Authors

[Jérôme Dokic](#)

dokic@ehess.fr

Institute Jean Nicod
Paris, France

[Margherita Arcangeli](#)

margheritarcangeli@gmail.com

Institute Jean Nicod
Paris, France

Commentator

[Anne-Sophie Brügger](#)

anne-sophie.bruegger@gmx.net

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

instance, we can visually imagine only what can be seen and auditorily imagine only what can be heard. To capture the latter examples, philosophers often talk of vision-like and audition-like imagination. More generally, the relevant type of imagination is *experience-like* or (as we shall also say) *experiential*, whether or not one believes that experiential imagination exhausts the field of possible imaginings.

However, the precise sense in which imagination is experiential remains a deep and complicated issue. In this essay, we would like

to inquire into the scope of experiential imagination. In particular, we want to relate the notion of experiential imagination to two important distinctions present in the contemporary literature on imagination, namely the distinction between sensory and cognitive imagination (Currie & Ravenscroft 2002; McGinn 2004) and the distinction between subjective and objective imagination (Vendler 1984; Dokic 2008). We aim at proposing, eventually, a systematic and hopefully enlightening taxonomy of the varieties of experiential imagination.

The essay is structured as follows: Section 2 tackles the broad phenomenological sense in which our imaginings are experiential. Sensory imagination will emerge as an important sub-type of experiential imagination.

Section 3 individuates two more fundamental sub-species of experiential imagination, namely objective and subjective imagination. We shall point out that this distinction maps onto an independently motivated distinction in the field of non-imaginative mental states, namely that between external and internal experiences. While external experiences (such as vision) are only accidentally *de se*, internal experiences (such as proprioception or agentive experience) are essentially or at least normally *de se*. The upshot will be that sensory imagination is best seen as a paradigmatic case of objective imagination.

Section 4 discusses the distinction between objective and subjective imagination, as Zeno Vendler introduces it on the basis of intuitive contrast examples. We shall show that Vendler's distinction diverges from ours, since it seems to hinge on a distinction between two ways the self can be involved in our imaginings. We shall suggest that the latter distinction is in fact orthogonal to our distinction between objective and subjective imagination (section 4.1). Moreover, upon closer look, the contrast examples offered by Vendler motivate our construction of the objective versus subjective distinction, which will prove to be more fruitful for the theory of imagination (section 4.2).

Section 5 presents the notion of cognitive or belief-like imagination and gives some reason to resist its interpretation as a form of non-ex-

periential imagination. Cognitive imagination can be construed as experiential, provided that at least some of our occurrent beliefs are conscious. Moreover, if belief is an experience, it is clearly an external experience. Therefore, cognitive imagination will emerge as a sub-species of objective imagination, along with sensory imagination.

Section 6 further investigates the domain of subjective imagination and its heterogeneity. We shall suggest that, along with proprioception, agentive experience, introspection, and feeling pain, subjective imagination may re-create other internal ways of gaining information about one's mental states, including beliefs.

Although much of our discussion in this essay belongs to conceptual clarification informed by phenomenological considerations, section 7 briefly describes several upshots of our account with respect to modal epistemology, cognitive resonance phenomena, mindreading, and imaginative identification. It is our contention that the relevance of the conceptual distinctions proposed by our taxonomy of experiential imagination has been crucially neglected in many important philosophical and scientific applications of the notion of imagination.

2 Experiential and sensory imagination

Let us start with Christopher Peacocke's analysis of imagination, which can help us to delineate what we mean by "experiential imagination." Peacocke (1985) puts forward what he calls the "General Hypothesis" about imagination, or GH (General Hypothesis) for short:

GH =_{DF} To imagine something is always at least to imagine, from the inside, being in some conscious state (Peacocke 1985, p. 21).

Peacocke does not offer an explicit definition of the phrase "from the inside", but we shall follow Kendall Walton's interpretation and assume that "the question of whether an imagining is from the inside arises only when what is imagined is an experience (broadly construed)" (Walton 1990, p. 31). For instance, I may ima-

gine being a descendant of Napoleon, but, according to Walton, my imagining does not essentially involve the perspective of any experience properly speaking. There is nothing it is like to be a descendant of Napoleon.² So there is no question of imagining “from the inside” having this relational property. In contrast, when I visually imagine a white sandy beach, my imagining involves an experiential perspective. I imagine “from the inside” a specific visual experience.

Peacocke’s notion of imagining from the inside is broadly related to other notions in the philosophical literature on imagination. For instance, Gregory Currie and Ian Ravenscroft introduce the notion of *recreative imagination* as the capacity to have “states that are not perceptions or beliefs or decisions or experiences of movements of one’s body, but which are in various ways like those states—like them in ways that enable the states possessed through imagination to mimic and, relative to certain purposes, to substitute for perceptions, beliefs, decisions, and experiences of movements” (Currie & Ravenscroft 2002, p. 12). Similarly, Alvin Goldman puts forward the notion of *enactment imagination* (or *E-imagination*) as “a matter of creating or trying to create in one’s own mind a selected mental state, or at least a rough facsimile of such a state, through the faculty of the imagination” (Goldman 2006, p. 42).³

Many other philosophers have held the view that imagination is the capacity to “modify” non-imaginative kinds of mental state (Husserl 1901; Meinong 1902; Mulligan 1999; Weinberg & Meskin 2006a, 2006b), where the relevant modification is to be understood as the “preservation” of some features of the non-imaginative states, such as part of their functional roles, despite phenomenological discrepancies or different overall cognitive underpinnings. This view is independent of a strong kind of simulationism, according to which each of several types of imagination shares with a proper non-imaginative *counterpart* some cognitive mechan-

ism (or set of information-processing systems), which is redeployed off-line.

To recapitulate, according to the terminology used in this essay, imagination is the general capacity to produce *sui generis* occurrent mental states, which we call “imaginings”. Whenever a subject imagines something, she is in a particular mental state of imagining. What type of mental state the subject is in depends on the non-imaginative conscious state that is re-created. Here we want to remain as neutral as possible with respect to the relationship between imaginings and their analogues in the non-imaginative mental realm. It is enough for our purposes to accept the idea that a phenomenologically useful taxonomy of imagination can be guided by a corresponding taxonomy of non-imaginative mental states (and perhaps also the other way around, as we shall suggest toward the end of the essay).

From now on, instead of using Peacocke’s phrase “imagining from the inside”, which is potentially misleading (see footnote 16 below), we are going to use phrases of the form “X-like imagination”, or “re-creating X” in imagination, where X is a type of non-imaginative state (as in “vision-like imagination”, or “re-creating a proprioceptive experience”). However, our use of these phrases should not be interpreted as carrying all the commitments of simulationist or recreative theories of imagination (whence the presence of the hyphen in “re-creating”).

GH turns out to be a general definition of imagination as essentially involving the perspective of a conscious experience—precisely what we call “experiential imagination”. Peacocke then introduces a more specific hypothesis precisely in order to identify sensory imagination as a sub-domain of experiential imagination.⁴ He himself calls this hypothesis the “Experiential Hypothesis”,

4 What is the relationship between sensory imagination and mental imagery? The latter phenomenon is at the heart of the well-known debate about the format of representations involved in cognitive tasks such as mental rotation (see Kosslyn 1980, 1994; Tye 1991; Pylyshyn 2002; Kosslyn et al. 2006). This debate concerns the kind of *content* of the relevant representations, and one of the issues is whether such content is propositional or iconic. In contrast, the notion of sensory imagination is defined here by reference to the psychological *mode* of the re-created mental state, namely a conscious perceptual experience. For our purposes we can leave open the nature of the contents of sensory imaginings.

² Throughout the paper, we assume that experiences are conscious mental states.

³ Goldman himself acknowledges that these two treatments of imagination are similar (Goldman 2006, p. 52, fn. 21).

but in order to avoid confusion and make it clear that only sensory imagination is at stake, we are going to call it the “Sensory Hypothesis” (or Sensory Hypothesis (SensH) for short), and rephrase it as follows:

SensH =_{Df} To imagine something sensorily is always at least to re-create some sensory experience.

For instance, imagining being in front of the Panthéon or at the helm of a yacht (Peacocke’s examples) may involve re-creating some visual experience as of being in front of the Panthéon or at the helm of the yacht.

Sensory imagination is not confined to vision. In Peacocke’s words, SensH deals with “imaginings describable pre-theoretically as visualizations, hearings in one’s head, or their analogues in other modalities” (Peacocke 1985, p. 22).⁵ A similar definition of sensory imagination can be found in the work of other philosophers (Kind 2001; Noordhof 2002; McGinn 2004). The same type of imagination has also been labeled “perception-like” (Currie & Ravenscroft 2002), “perceptual imagination” (Chalmers 2002), and even “experiential imagination” (Carruthers 2002).

To the extent that SensH is concerned only with cases in which the subject re-creates a specific type of experience, namely sensory experience, it deals with a sub-type of experiential imagination as covered by GH, namely sensory imagination. At this point, the question arises as to what other types of experiential imagination there are beyond the sensory type. Walton suggests that the notion of experience at stake in GH should be interpreted in a broad way, and we may wonder about its precise breadth.

3 Objective and subjective imagination

Peacocke himself intends GH to cover genuine instances of experiential imagination that are not covered by SensH—what we shall call non-sensory

imagination. For instance, one can imagine “the conscious, subjective components of intentional action” (Peacocke 1985, p. 22). On Peacocke’s view, imagining playing the Waldstein sonata may involve re-creating a non-sensory experience, namely the intimate experience one has of one’s own action while or in acting.

Of course, the precise nature of what we may call “motor imagery” is controversial.⁶ Currie and Ravenscroft suggest that “motor images have as their counterparts perceptions of bodily movements. They have as their contents active movements of one’s body” (Currie & Ravenscroft 2002, p. 88). So on Currie and Ravenscroft’s suggestion, imagining playing the Waldstein sonata involves re-creating the perception of bodily movements.

Certainly, in order to imagine performing an action, it is not enough to re-create a *visual* experience of the appropriate bodily movements—otherwise, the relevant type of imagining would belong to sensory imagination after all. Alternatively, one might suggest that motor imagery involves re-creating a *proprioceptive* experience of the appropriate bodily movements. However, such imagining does not entail re-creating an agential experience, even if it may accompany the latter. In a similar vein, Goldman claims that motor imagery “is the representation or imagination of executing bodily movement” and has as its counterpart “events of motor production, events occurring in the motor cortex that direct behavior” (Goldman 2006, pp. 157–158). Following Goldman, we can say that imagining playing the Waldstein sonata may involve re-creating an execution of the appropriate bodily movements.

In fact, an ordinary case of imagining playing the Waldstein will probably involve (at least) three types of imagining:

- Imagining *seeing* movements of one’s fingers on the keyboard.
- Imagining *having a proprioceptive experience* of these movements.
- Imagining *playing* the sonata.

⁵ At this point we can count at least the five senses (vision, audition, touch, taste, and olfaction) as sensory modalities. Later on, we shall suggest that a sensory modality involves an *external* perceptual perspective on the world. This excludes proprioception and the sense of agency as sensory modalities, insofar as they involve *internal* perspectives on oneself.

⁶ See section 7.2. As Thomas Metzinger reminded us, the existence of motor imagery has been acknowledged by twentieth century phenomenology. For instance, Karl Jaspers has coined the German term “Vollzugsbewusstsein”, which can be translated as “executive consciousness”.

The three types of imagining are typically entangled within a single imaginative endeavor. That is, someone who imagines playing the sonata will typically imagine having a proprioceptive experience of her fingers running on the keyboard but also various sensory experiences: visual experiences of her moving fingers and auditory experiences of the music. Still, each type is essentially distinct from the others, and might even be dissociable in special circumstances (although we do not want to insist too much on the possibility of such dissociation). Suppose for instance that one imagines one's limbs being remotely controlled. One can imagine from a proprioceptive perspective one's arms and legs going through the motions characteristic of playing the piano without imagining oneself playing the piano. In this case, (ii) is instantiated but (iii) is not. More controversially, suppose that one imagines oneself being selectively anesthetized, or in the situation of a deafferented subject.⁷ Perhaps one can then imagine playing the piano without imagining having a proprioceptive experience; (iii) but not (ii) would be instantiated. Given the role of proprioceptive feedback in the ordinary execution of action, it is probably hard if not impossible to imagine playing a whole sonata in the absence of any proprioceptive-like imagining, but the relevant dissociation is in principle possible for simpler actions, such as stretching one's finger. Finally, it seems possible to imagine playing the piano without re-creating any visual or auditory experience. For instance, one can imagine playing the sonata with one's eyes closed or one's ears blocked. Here, (iii) is instantiated but (i) is not. Again, given the role of sensory feedback in the ordinary execution of action, it might be hard to form such a selective imagining, especially if the action gets complicated.

The upshot of the foregoing discussion is that only (iii) is a genuine case of motor imagery. It involves the re-creation of what philosophers of action call the "sense of agency" or the "sense of control" (see e.g., Haggard 2005 and Pacherie 2007). Since the sense of agency or control is a conscious experience, motor imagery clearly falls under the umbrella of experiential imagination.

⁷ Deafferented patients have lost the sense of proprioception; see e.g., Cole (1995) and Gallagher (2005).

Moreover, to the extent that motor imagery is (at least in principle) dissociable from sensory imagination, even if it typically depends on the latter, it is a case of non-sensory imagination.⁸

What about (ii)? Proprioception is arguably a mode of perception; it is a way of perceiving the spatial disposition of one's body.⁹ In this respect, (ii) is like (i), which is a case of sensory imagination. However, proprioception is also essentially or at least normally a way of gaining information about oneself; what proprioception is about is a bodily state of oneself. In this respect, (ii) is more like (iii), which also involves a way of gaining information about oneself, and more precisely one's actions.¹⁰

What unifies (ii) and (iii) as cases of non-sensory imagination is the fact that what is re-created is a (non-imaginative) *internal* experience. An internal experience is essentially or at least normally *de se*, in the following sense: it is supposed to be about a mental or bodily state of oneself. Proprioceptive and agentive experiences are both internal in this sense. At least in normal circumstances, one cannot have a proprioceptive experience of another's body or a sense of agency for another's action. In contrast, all cases of sensory imagination are such that what is re-created is a (non-imaginative) *external* experience. An external experience is typically about the external world and is only accidentally *de se*. For instance, vision is an

⁸ Even if it turns out that motor imagery is constitutively dependent on sensory imagination, it is clearly not fully sensory, as we will shortly show. Note also that if motor imagery can be conceived as the re-creation of an essentially active phenomenon, namely the sense of agency or control, it need not be itself active. Although we cannot dwell on this issue here, imaginings can be either active, when we deliberately imagine something, or passive, as for instance when we are lost in an episode of mind wandering (see footnote 22).

⁹ If proprioception is a case of perception, there must be proprioceptive experiences. This has been contested, especially by Anscombe (1957). However, in our view, Anscombe conflates two different claims. The first claim, which we accept, is that there are no proprioceptive *sensations*. Proprioception is not a case of sensory perception. The second claim, which we reject, is that proprioception does not involve any conscious *experience*. Even if there are no proprioceptive sensations, we are consciously aware of the positions and movements of our body.

¹⁰ The idea that there are "self-informative methods," i.e., ways of finding out about oneself, is pervasive in John Perry's theory of self-knowledge; for a recent statement, see Perry (2011). As Perry makes clear, these methods can be either metaphysically or merely architecturally guaranteed. François Recanati makes use of a similar idea in his account of perspectival thought (Recanati 2007) and mental files (Recanati 2012); for instance, he writes: "In virtue of being a certain individual, I am in a position to gain information concerning that individual in all sorts of ways in which I can gain information about no one else, e.g. through proprioception and kinaesthesia" (Recanati 2007, p. 262).

Objective imagination	Subjective imagination	
Sensory imagination	Proprioception-like imagination	Action-like imagination
<p>I re-create in imagination a visual experience of <i>my fingers running on the keyboard</i></p> <p>I re-create in imagination an auditory experience of <i>music</i></p> <p>I re-create in imagination a multimodal experience of <i>the music as caused by the motions of my fingers</i></p>	<p>I re-create in imagination a proprioceptive experience of <i>my fingers running on the keyboard</i></p>	<p>I re-create in imagination the action of <i>running my fingers on the keyboard</i></p> <p>I re-create in imagination the action of <i>playing the sonata</i></p>

Figure 1: Types of imagining involved in playing a sonata

external experience; it is a way of gaining information about one's immediate surroundings, whether or not one also sees oneself.¹¹

These considerations allow us to give a more fine-grained analysis of the realm of experiential imagination based on the external versus internal contrast, rather than the sensory versus non-sensory contrast. In a nutshell, we

¹¹ This is an oversimplification, since many ordinary experiences have presumably both internal and external aspects. On the one hand, vision might involve both exteroception and interoception (Gibson 1966). On the other hand, proprioception and other forms of bodily experience often rely on visual information (Botvinick & Cohen 1998; de Vignemont 2013). Still, the external aspect of many ordinary visual experiences is clearly dominant, while visually aided proprioception remains essentially a way of gaining information about oneself, and thus is an internal experience in our sense.

can say that experiential imagination comes in two varieties. Experiential imagination can re-create: (a) some external experience—e.g., a way of gaining information about the world (e.g., I imagine *seeing* Superman flying in the air), and (b) some internal experience—e.g., a way of gaining information about oneself (e.g., I imagine *having a proprioceptive experience of* flying in the air). Following Jérôme Dokic (2008), we shall call (a) “objective imagination” and (b) “subjective imagination”; see figure 1.¹²

¹² To make our terminology as clear as possible, the distinction between internal and external experiences concerns the realm of non-imaginative states, while the analogous distinction between subjective and objective imagination concerns the realm of imaginative

We may thus introduce two other hypotheses subordinate to [GH](#), which we call the Objective Hypothesis (ObjH) and the Subjective Hypothesis (SubjH):

ObjH =_{Df} To imagine something objectively is always at least to re-create some external experience.

SubjH =_{Df} To imagine something subjectively is always at least to re-create some internal experience.

Sensory imagination forms an important subclass of experiential imagination, but it can also be seen as a paradigmatic case of objective imagination, since it involves re-creating an *external* experience. Experiential imagination is not merely objective imagination, since another sub-class of experiential imagination, namely subjective imagination, is constituted by cases in which an *internal* experience is re-created. For instance, imagining having one's legs crossed or driving a Ferrari may involve re-creating some internal non-sensory experience, namely a proprioceptive and/or agentive experience as of having one's legs crossed or driving a Ferrari.

To sum up, we have identified two important varieties of imagination that seem to exhaust the domain of experiential imagination: objective and subjective imagination.¹³ We have argued that this distinction, which gives rise to phenomenologically different imaginings, traces back to an independent distinction within the domain of non-imaginative experiences, between external and internal experiences. We have also claimed that sensory imagination, which is the variety of experiential imagination most commonly recognized, should be seen as a paradigmatic example of objective imagination. More should be said about the distinction between objective and subjective imagination. For instance, questions arise as to whether sensory imagination exhausts the field of objective imagination and as to whether

subjective imagination encompasses more than proprioceptive or agentive experiences.

The remainder of the paper is devoted to further clarification of the notions of objective and subjective imagination. We shall begin with a comparison between our own proposal and Zeno Vendler's observations about imagination.

4 Vendler's varieties of imagination

A well-informed reader might think that our distinction between objective and subjective imagination is the same as a homonymous distinction introduced by [Vendler \(1984\)](#). Certainly Vendler intends to capture two phenomenologically different ways of imagining, which potentially correspond to our distinction between external and internal experiential perspectives (perspectives on the world and perspectives on oneself). However, he also gives a *prima facie* interpretation of the distinction between objective and subjective imagination, which has more to do with the way the self is involved in our imaginings than with the distinction between external and internal experiences. On this interpretation, Vendler's notions of objective and subjective imagination arguably diverge from ours. Let us start with Vendler's interpretation of these notions (section 4.1) and then move to a deeper analysis of the contrast examples offered by Vendler in order to motivate his distinction (section 4.2). In so doing, we shall show that our construction of the objective versus subjective distinction is more helpful in order to map the realm of experiential imagination.

4.1 Two kinds of self-involvement

[Vendler \(1984\)](#) suggests that the phrase "S imagines doing A" invites what he calls "subjective" imagination, while the phrase "S imagines herself/himself doing A" can be used to describe "objective" imagination. *Prima facie*, Vendler seems to interpret the distinction between subjective and objective imagination in terms of two ways in which the self can be involved in our imaginings—implicitly or explicitly.

Subjective imagination concerns cases in which the self is implicitly involved in the imagining, whereas objective imagination concerns

states. The question of whether imaginings themselves can be said to be internal or external is not raised in this essay.

¹³ Note that our definition leaves open the possibility that a particular imagining is both objective and subjective, to the extent that the re-created experience has both external and internal aspects (see footnote 11).

cases in which the self is explicitly involved in the imagining. This is why the phrase “imagining doing A”, which does not explicitly mention the agent of the action A, is best used to describe subjective imagination, whereas the phrase “imagining *myself* doing A”, which explicitly mentions myself as the agent of the action A, is more suitable to the description of objective imagination.

The self is implicitly involved in an imagining when it fixes the point of view internal to the imagined scene without being a constituent of that scene. One can imagine seeing the Panthéon from the other end of rue Soufflot without imagining oneself as another object in the scene. Still, the scene is imagined from a specific point of view, as defined by a virtual self. One can also imagine seeing oneself in front of the Panthéon. In such a case, the self is a constituent of the imagined scene—it is explicitly represented as a part of the imagining’s content.

Of course, when one imagines seeing oneself in front of the Panthéon, one’s imagining also involves the self implicitly. One imagines a scene from the perspective of a virtual self, which is distinct from oneself as a constituent of the scene. As a consequence, Vendler makes clear that subjective and objective imagination are not mutually exclusive. Commenting on Vendler’s distinction, François Recanati concurs, writing that “the objective imagination is a particular case of the subjective” (Recanati 2007, p. 196).

It should be sufficiently apparent that the distinction between implicit and explicit self-involvement is a matter of the imagining’s *content* and more precisely deals with the issue of how the self is involved in imagination. In contrast, the distinction between internal and external experiential perspectives has to do with the *mode* re-created in imagination, respectively an external and an internal experience. Therefore, the two distinctions answer different questions and turn out to be orthogonal.

First, objective imagination can involve the self either implicitly or explicitly (but not both at the same time). This is easily seen by considering the Panthéon example. Peacocke

himself suggests another relevant case. He observes that the phrase “imagining being seated on a horse” is ambiguous between adopting the point of view of the rider (namely oneself) and adopting the point of view of someone else who could see the rider (see Peacocke 1985, p. 23). If the relevant perspective is that of the rider (namely oneself), the self need not be a constituent of the imagined scene—in this case (where the rider does not see any part of her body), it is implicitly involved in the imagining. In contrast, if the relevant perspective embraces oneself as the rider, the self is explicitly involved; it figures in the content of the imagining. However, both interpretations involve *visual* (i.e., external) perspectives, so what is at stake is a distinction within *objective* imagination rather than a contrast between subjective and objective imagination.¹⁴

Second, it is at least arguable that subjective imagination can involve the self either implicitly or explicitly. Suppose that one subjectively imagines swimming in the ocean. One may re-create the internal experience of what Marc Jeannerod & Elisabeth Pacherie (2004) call a “naked” intention (in action), which precisely does not involve an explicit representation of the agent. In this case, no self is part of the representational content of one’s imagining. One subjectively imagines swimming without imagining the agent as such, whether oneself or anyone else. However, one might also re-create a more complex internal experience, whose content embraces oneself as the agent of the action of swimming. Accordingly, in this case, the self (oneself) is explicitly represented in the content of one’s subjective imagining. One subjectively imagines a particular agent swimming; in Vendler’s example, that particular agent is oneself.

One might object to the last point and claim that the self is never an object of internal experience. One can have at best internal experiences of particular mental states, such as inten-

¹⁴ One might object that both cases involve subjective imagination, since the visual perspective of the rider, even if she does not see her own body, is tied to her proprioceptive experience; see the *caveat* voiced in footnote 11 above. Again, it might be that the distinction between subjective and objective imagination has really to do with the distinction between re-creating predominantly internal and re-creating predominantly external experiences.

Experiential Imagination		
	Explicit self-involvement	Implicit self-involvement
Subjective	I re-create in imagination a proprioceptive and/or agentive experience of <i>myself flying in the air</i>	I re-create in imagination a proprioceptive and/or agentive experience of <i>flying in the air</i>
Objective	I re-create in imagination a visual experience of <i>myself being seated on a horse</i>	I re-create in imagination a visual experience of <i>being seated on a horse</i>

Figure 2: Explicit and implicit self-involvement in subjective and objective imagination

tions in action, but never of oneself having those mental states. However, this is a substantial claim that certainly needs to be backed up by careful arguments. Note that the assumption that the self can figure in the content of an internal experience is in principle compatible with the Humean point that the self cannot be introspected. Introspection, conceived as a form of inner perception, is only one type of internal experience. Perhaps there are non-introspective cases of internal experience whose explicit contents cannot be fully specified except by using the first-person pronoun. For instance, one might argue that at least some cases of proprioception as well as internal experiences of controlling one's body as a whole give us access to one's self, or at least to the boundaries between oneself and the rest of the environment.¹⁵

Consider other examples offered by Vendler. When you imagine yourself eating a lemon by imagining your pinched face, your imagining is

explicitly self-involving and might be fulfilled via objective imagination, such as visual imagination, but also via subjective imagination, such as proprioceptive imagination, at least to the extent that it recreates an internal experience of your bodily self. What about imagining implicitly involving the self? If while imagining eating a lemon, the subject imagines the action of eating a lemon and nothing else, she is exploiting her subjective imagination, insofar as she is re-creating an agentive perspective. It seems possible to imagine eating a lemon via objective imagination too, for instance by re-creating a visual experience as of an action independently of any identification of the agent.

To sum up, while the distinction between subjective and objective imagination seems to capture two forms or modalities of imagination, the distinction between two kinds of self-involvement, although important in itself, is less relevant to a taxonomy of experiential imagination. The orthogonality of these distinctions is shown again in figure 2. In the following sub-

¹⁵ For relevant discussion, see e.g., Cassam (1999), Bermúdez et al. (1995), Bermúdez (1998), Metzinger (2003), and Peacocke (2014).

section we shall further motivate our hypothesis that Vendler's own contrast examples are best understood in terms of the independently motivated distinction between internal and external experiences.

4.2 Vendler's examples revisited

Aside from his interpretation of subjective imagination as implicitly self-involving and objective imagination as explicitly self-involving, Vendler clearly draws our attention to two ways of imagining a given action, which have quite different phenomenological profiles. In his own words:

We are looking down upon the ocean from a cliff. The water is rough and cold, yet there are some swimmers riding the waves. 'Just imagine swimming in that water' says my friend, and I know what to do. 'Brr!', I say as I imagine the cold, the salty taste, the tug of the current, and so forth. Had he said 'Just imagine yourself swimming in that water', I could comply in another way, too: by picturing myself being tossed about, a scrawny body bobbing up and down in the foamy waste. (Vendler 1984, p. 43)

As some of Vendler's other examples show, the relevant distinction is not restricted to imagining actions:

In order to familiarize yourselves with this distinction, imagine eating a lemon (sour taste), and then imagine yourself eating a lemon (pinched face); imagine being on the rack (agony), and then yourself being on the rack (distorted limbs); imagine whistling in the dark (sensation of puckered lips), and then yourself whistling in the dark (distance uncertain, but coming closer); and so forth. (Vendler 1984, p. 43)

It is not immediately clear what is common to all cases of subjective or objective imagination in Vendler's examples. Consider the suggestion

that the relevant distinction can be explained at the level of the *states* represented by the imaginings. Subjective imagination would involve imagining states that *cannot* be imagined objectively. For instance, in imagining swimming in the water, I also imagine proprioceptive experiences, which (one might argue) cannot be imagined objectively. How could we *visually* imagine such experiences, which are essentially *felt*?

However, it is not obvious that the essence of the distinction between subjective and objective imagination can be fully captured by reference to the imagined states. One can imagine having one's legs crossed via subjective imagination, but also via objective imagination. The first type of imagining is akin to proprioception (one imagines feeling one's legs crossed), while the second type of imagining is akin to vision (one visualizes oneself with one's legs crossed). Yet these imaginings are about the same bodily condition—having one's legs crossed.

Similarly, the very same action of swimming in the ocean can be imagined subjectively or objectively. The case of pain is more controversial, but if one can be visually aware that someone is in pain (by observing pain-related behavior), then one can imagine the very same pain state either subjectively or objectively. The difference between the relevant imaginings must lie elsewhere.

We are now in the position to see that we were on the right track and that Vendler's contrast examples are plausibly construed as involving different experiential perspectives on a given scene, either internal (perspectives on oneself) or external (perspectives on the world). Subjective imagination has to do with the former, and objective imagination with the latter. This is easily seen by considering the example of imagining whistling in the dark. Vendler contrasts the subjective case, in which the subject imagines the sensation of puckered lips, with the objective case, in which the subject imagines the distance uncertain, but coming closer. In other words, what Vendler seems to contrast is proprioception-like imagination with auditory imagination or, in our terminology, an internal experiential perspective with an external one.

More generally, Vendler seems to be concerned with the difference between, on the one hand, imagining doing an action (e.g., swimming, eating, whistling, etc.) or having pain (e.g., agony), where what the imaginer re-creates is the relevant experience and, on the other hand, imagining pieces of behaviour that reveal the very same experience (e.g., visualizing an eating mouth or a body in agony), where what the imaginer re-creates is an external perspective on the relevant experience.¹⁶

Let us note that, in order to make his contrast more realistic, Vendler gives us complex examples, where more than one experience is involved. So for instance, his example of imagining swimming in the ocean clearly belongs to subjective imagination, since the re-creation of a proprioceptive and/or agentive experience is involved. As Vendler suggests, though, when you fulfill this imagining you can also re-create various external experiences, such as “the cold, the salty taste, the tug of the current, and so forth”. The same is true in the case of imagining eating a lemon. When you imagine eating a lemon, you re-create in imagination an internal experience (e.g., the proprioceptive and/or agentive experience of eating), but your imagining can be accompanied by others that re-create external experiences (e.g., the sour taste, the yellow lemon).

The discussion of Vendler’s distinction has led us to strengthen our taxonomy of experiential imagination. So far we have seen that, first, all cases covered by *SenH* seem to be cases of objective imagination (and thus covered by *ObjH*), which involves re-creating some external experience. Second, all cases covered by *SubjH* arguably involve re-creating some internal experience.

However, another important type of imagination emerges from the literature on imagin-

ation, namely cognitive imagination, which has been defined as belief-like and typically contrasted with sensory or even experiential imagination.

5 Cognitive imagination

Many authors contrast sensory imagination with cognitive imagination (“imagining that,” or “propositional imagination”), which has been defined as belief-like (Mulligan 1999; Currie & Ravenscroft 2002; McGinn 2004; Goldman 2006; Weinberg & Meskin 2006b; Arcangeli 2011a).¹⁷ Cognitive imagination seems to be relatively autonomous from sensory imagination. For instance, one can imagine that poverty has been reduced in the world independently of re-creating any visual, auditory, tactile, etc., experience. Of course the autonomy of cognitive imagination relative to sensory imagination echoes the autonomy of belief relative to sensory perception (one can believe that poverty must be reduced in the world without perceiving anything).

Cognitive imagination is by essence non-sensory, but given our previous discussion, it does not exhaust the field of non-sensory imaginings. Re-creating in imagination some internal experience is presumably non-cognitive (in the relevant sense of being belief-like), but it is non-sensory as well. Thus we have, at least *prima facie*, three types of potentially dissociable imagination: sensory non-cognitive imagination (e.g., I imagine hearing a piece of music, such as Ravel’s *Concerto pour la main gauche*), non-sensory non-cognitive imagination (e.g., I imagine having the proprioceptive experience of being one-armed), and non-sensory cognitive imagination (e.g., I imagine that Maurice Ravel has created a piano piece especially for me).

One might argue that cognitive imagination is not only non-sensory but non-experiential as

¹⁶ A similar point is made by Mike Martin when he draws a distinction between “cases in which there is just an itch in the left thigh” in imagination and cases “in which one imagines some person whose behaviour reveals that they have an itch” (Martin 2002, p. 406, fn. 35; see also Dorsch 2012). However, according to his terminology, only the former cases count as being “from the inside”. Very often in the literature, the phrase “imagining from the inside” is used in this narrow sense (to refer to subjective imagination in our terminology) more than the broad sense meant by Peacocke (which refers to experiential imagination as a whole).

¹⁷ In fact, Mulligan speaks of a judgement-like, rather than a belief-like, type of imagination, which he calls “supposition”. It is not entirely clear whether his notion of supposition can be equated with what we call “cognitive imagination.” Very often in the literature, supposition is taken to be belief-like and, as such, nothing but cognitive imagination (Nichols & Stich 2003; McGinn 2004; Goldman 2006). An alternative view is that supposition is a *sui generis* type of imagination akin to *acceptance* rather than belief (Arcangeli 2011b). However, for present purposes we will skip this issue and consider only belief-like imagination.

well and as such lies outside the scope of [GH](#). According to a standard view, beliefs, even if they can be occurrent, are not conscious experiences strictly speaking. On this view, an occurrent belief may be accompanied by various experiences (mental images, feelings, emotions, etc.), but there is nothing it is like to have a belief.¹⁸ Now this view has recently come under attack by philosophers who acknowledge the existence of a doxastic phenomenology, i.e., a kind of phenomenology characteristic of belief (see the debates on cognitive phenomenology in [Bayne & Montague 2011](#)). On this alternative view, there is something it is like to have an occurrent belief, which is reducible to neither sensory nor affective phenomenology. At least some occurrent beliefs would be *sui generis* conscious experiences.¹⁹

If the alternative view is broadly correct, some beliefs lie within the scope of [GH](#).²⁰ In order to capture cognitive imagination as a putative form of experiential imagination, let us introduce another specific hypothesis subordinate to [GH](#), which we call the Cognitive Hypothesis (C):

CogH =_{Df} To imagine something cognitively is always at least to re-create a conscious occurrent belief.

For instance, cognitively imagining that quantum physics is false or that this pen is an alien involves re-creating the conscious occurrent belief that quantum physics is false or that this pen is an alien. In general, one may surmise that anything that can be consciously believed can be cognitively imagined.

We have suggested that experiential imagination divides into two sub-domains only, namely subjective and objective imagination (covered by [SubjH](#) and [ObjH](#), respectively). In addition, sensory imagination (covered by [SensH](#))

emerged as a species of objective imagination and non-sensory non-cognitive types of imagination (e.g., proprioception-like and agentive-like imagination) have been described as paradigmatic cases of subjective imagination. What about cognitive imagination (covered by [CogH](#))? Is it a type of objective or of subjective imagination? Or should we acknowledge a third class of experiential imaginings that are neither objective nor subjective?²¹

We have introduced the distinction between objective and subjective imagination as the imaginative analogue of the distinction between external and internal experience. As we have seen, many external experiences are ways of gaining information about the world, and many internal experiences are ways of gaining information about oneself. Now one might claim that belief, unlike perceptual or introspective experience, is not individuated in terms of ways of gaining information. Of course some of our beliefs result from various ways of gaining information about the world and ourselves, but it is logically possible to have a belief that is not the result of any source of information. Does it follow that belief as an experience is neither external nor internal? Not really, for an external experience has been more fundamentally defined as being accidentally *de se*, whereas an internal experience is essentially or at least normally *de se*. In this more fundamental sense, if belief is an experience, it is clearly an external experience: one can believe all sorts of states of affairs that do not involve or concern oneself. It follows that cognitive imagination, as the re-creation of an external doxastic experience, is better seen as a sub-species of objective imagination, along with sensory imagination. Objective imagination then emerges as a heterogeneous domain, but where at least two clearly different types of imagining can be distinguished (see figure 3).

¹⁸ See [Metzinger \(2003\)](#), [Tye \(2009\)](#), and [Carruthers & Veillet \(2011\)](#). Note that the standard view can lead to different attitudes toward the notion of cognitive imagination. On one attitude, cognitive imagination exists but is non-experiential. On another attitude, cognitive imagination does not exist or wholly reduces to sensory imagination (if, for instance, it is construed as auditory verbal imagination).

¹⁹ [Crane \(2013\)](#) defends a closely related view, according to which episodes of thinking, although not beliefs themselves, are phenomenally conscious. [CogH](#) can easily be adapted to accommodate Crane's view.

²⁰ In conversation, Peacocke confirmed that he intends [GH](#) to cover at least some cases of belief-like imagination.

²¹ Moreover, the question of whether these varieties of imagination exhaust the field of experiential imagination remains open. In order to answer it we would have to inquire as to whether there are other types of imagination, such as desiderative or desire-like imagination (see [Currie & Ravenscroft 2002](#) and [Doggett & Egan 2007](#) for a positive view, and [Weinberg & Meskin 2006a](#) and [Kind 2011](#) for a critical view), affective or emotion-like imagination (see [Goldman 2006](#) for a positive view, and [Currie & Ravenscroft 2002](#) for a critical view) and judgement-like or acceptance-like imagination (see footnote 17). For lack of space, we have to defer this inquiry to another occasion.

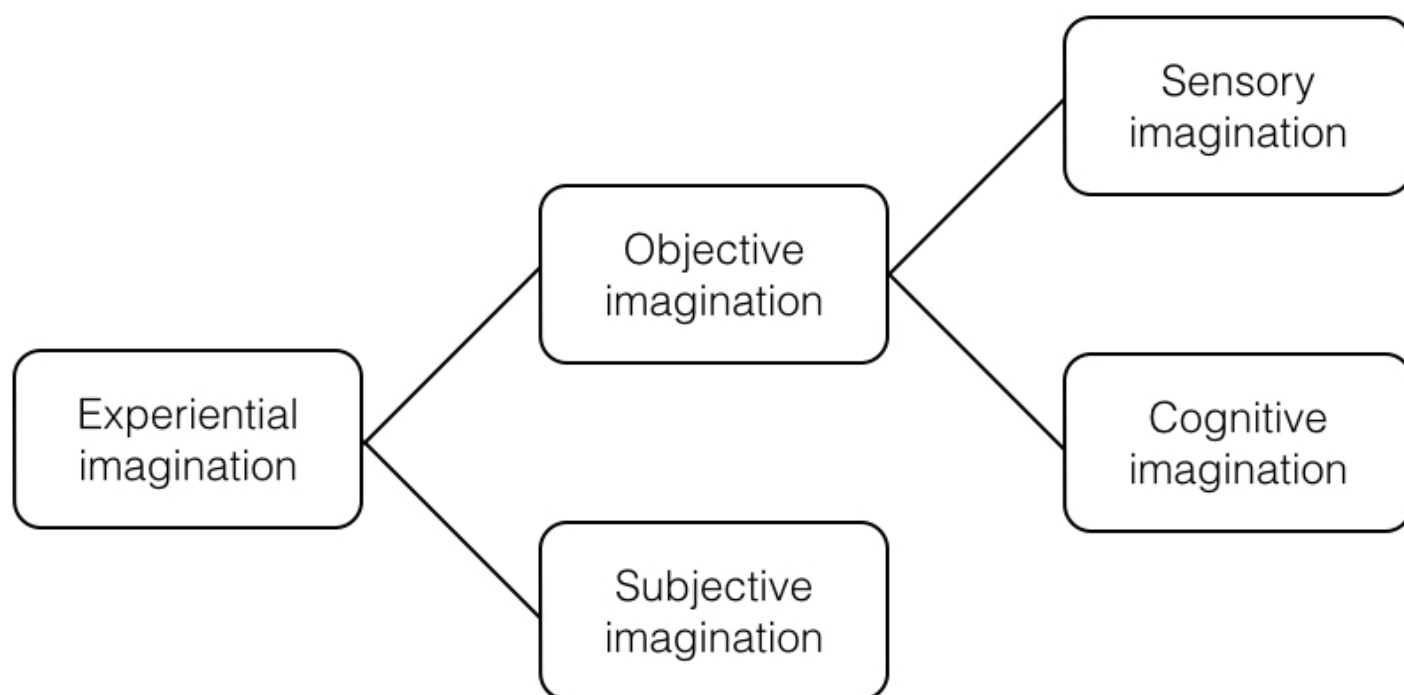


Figure 3: The varieties of experiential imagination

6 The scope of subjective imagination

As we have seen, subjective imagination involves re-creating various ways of gaining information about ourselves, such as proprioceptive or agentive experience. Now we also have ways of gaining information about our own sensory experiences, as well as about our own beliefs. We seem to be able to form self-ascriptions of the form “I see x ” or “I believe that p ” without relying on independent background beliefs. The nature of this ability is controversial. Some philosophers claim that both sensory experiences and beliefs can be introspected (e.g., [Goldman 2006](#)). Thus, we should be open to the possibility of re-creating in imagination an introspective experience of a visual experience or an occurrent belief. Other philosophers reject the notion of introspection altogether and consider that self-ascription of sensory experience or belief can follow a purely theoretical procedure known as an “ascent routine” (see [Evans 1982](#) and [Gordon 1995](#) for the case of belief, and [Byrne 2010](#) for suggestions about how to extend the ascent routine to sensory experience).

The question arises as to what types of internal experience can be re-created in imagination, i.e., what the scope of subjective imagination is. In a sense, this question is hostage to an

independent theory of internal experience, appropriate to sensory experience or belief. Obviously, we cannot settle the matter in this exploratory essay. Still, before moving to the penultimate section, we would like to suggest that phenomenologically accessible distinctions within the realm of imagination might be conceived as (usually neglected) constraints on a correct theory of internal experience. We shall focus on belief, but similar observations can be made for the case of sensory experience.

There is some phenomenological evidence that subjective imagination can capture an internal perspective on at least some beliefs. Consider an atheist who tries to imagine what it is like to believe in God. One might argue that this involves re-creating some internal experience of an occurrent belief in God. At least the atheist’s imagining seems different from two other types of imagining, namely imagining believing in God and imagining believing that one believes in God.

First, it is different from re-creating in imagination an occurrent belief in God, which would be an example of cognitive imagination. The latter imagining does not have belief as a constituent of its content; one cognitively imagines God himself, rather than some belief in

his existence. In general, an imagining that re-creates the non-imaginative state *M* need not have *M* as part of its content; the imagining itself is an imaginative re-creation of *M*, but it is not about *M* (Currie & Ravenscroft 2002, p. 27; see also Burge 2005, p. 63, for the corresponding point about sensory imagination). In contrast, the atheist's imagining essentially has belief as one of the constituents of its content; one imagines a particular belief in God. A related difference is that re-creating an occurrent belief in God is re-creating an external experience of a God-involving world. However, the atheist might want to imagine what it is like to believe in God without taking a stance on the presence of God in the imaginary world. Her imagining is focused on the belief in God, independently of whether it is true or false (even though as an atheist she believes it to be false).

Second, the atheist's imagining is different from re-creating an occurrent higher-order belief that one believes in God, i.e., imagining that one believes in God. Intuitively, the former imagining is more specific than the latter (which is another example of cognitive imagination). Imagining having the higher-order belief that one believes in God involves re-creating an external experience of one's belief in God. However, the atheist is not merely imagining that she or someone else has a belief in God. She wants to get into the believer's mind and re-create in imagination an *internal* perspective on some occurrent belief in God.

In the context of GH, the apparent existence of cases of subjective imagination where the re-created experience is an internal experience of belief can be seen as a constraint on a correct account of the way we gain information about our own beliefs. The introspective account can offer a straightforward explanation of the atheist's imagining as involving the re-creation of an introspective experience, as opposed to a mere higher-order belief, about the belief in God. *Prima facie*, the ascent-routine account has fewer resources to give justice to the relevant phenomenology. It might not be impossible to do so, though, if experiential imagination can also re-create complex cognitive processes such as going through an ascent routine. Again, we

have to leave the discussion for another occasion. It is enough for our purposes to gesture toward the possibility of extending the scope of subjective imagination to encompass more or less specific internal perspectives on beliefs, even if further argument is certainly needed.

7 Some applications

In this penultimate section, we would like to briefly illustrate how the fate of important claims about imagination made by philosophers and scientists depends on something like our taxonomy of experiential imagination. Although we believe that this taxonomy has philosophical value in its own right, we also would like to show that it is connected to central issues in philosophy and cognitive science. These issues concern, respectively, modal epistemology (section 7.1), cognitive resonance (section 7.2), mindreading (section 7.2), and imaginative identification (section 7.4). Our discussion in what follows, though, can only be rather programmatic in contrast to the rest of the essay.²²

7.1 Modal epistemology

Imagination has been traditionally construed as providing evidence for modal claims. For instance, many philosophers since Descartes have suggested that what can be imagined is metaphysically possible. On the other hand, imagination has been shown to produce various sorts of modal illusions (Kripke 1980; Gendler & Hawthorne 2002). The main challenge faced by proponents of an internal relation between imagination and possibility (perhaps via conceivability) is thus to distinguish proper and improper

²² This is only a selection of issues where we think our phenomenological and conceptual distinctions are relevant. We wish we had space to discuss other topics of relevance to the theory of imagination, such as mental time travel (Schacter & Addis 2007), dreams (Windt 2014), and mind wandering (Metzinger 2013). For instance, there are interesting issues having to do with the apparent lack of reflexivity of mind wandering episodes, and the tendency for the mind wanderer to identify herself with imagined protagonists (Metzinger 2013). A speculative hypothesis is that the passivity of mind wandering episodes causes various metacognitive errors, such as the error of confusing a case of subjective imagination with a genuine case of internal experience, which leads the imaginer to self-identify with the subject of the imagined mental state. Again, we have to leave this fascinating issue to another occasion.

uses of imagination, i.e., those uses that provide, and those that do not provide, evidence for modal claims. One might suggest, for instance, that proper uses of imagination require a certain format that other uses lack (Nichols 2006; Weinberg & Meskin 2006b).

In our view, there is an additional criterion that must be taken into account in these debates, which concerns the type of non-imaginative state that is re-created by the relevant imaginings. It might be that only some types of imagination are internally related to modal properties. For instance, it is not clear that cognitive imagination is essentially related to possibility. Assuming the correctness of our claim that cognitive imagination re-creates belief, the fact that one can cognitively imagine that p is no more evidence that p is possible than the mere fact that one believes that p . After all, one can believe all sorts of metaphysically impossible states of affairs (such as that Hesperus and Phosphorus are distinct celestial bodies).

The challenge is then to identify the types of imagination, if any, that are essentially or at least reliably related to what is metaphysically possible. One hypothesis, voiced by Dokic (2008), is to focus on types of imagination that re-create states of (actual or potential) knowledge. On this hypothesis, some uses of imagination are guides to possible contents because they are guides to the possibility of *knowing*. To the extent that sensory perception is commonly thought to be a source of knowledge, sensory imagination could be reliably linked to the possibility of what is imagined in this way (see also Williamson 2008).

This is not to say that cognitive imagination has no role to play in providing evidence for modal claims. Just as belief can be grounded on sensory perception and thereby be counted as knowledge, a single imagining might re-create not only belief and perception separately, but the complex mental state of believing that p on the basis of suitable sensory evidence (see Dokic 2008). The resulting imagining would be neither purely sensory nor purely cognitive, but to the extent that it re-creates a non-imaginative state of knowledge, its content might be bound to what is metaphysically possible.

7.2 Cognitive resonance

If we are right, there is a phenomenologically accessible distinction between objective and subjective imagination. What it is like to visually imagine an action or a painful experience is typically different from what it is like to subjectively imagine acting or having pain. However, this distinction is rarely made explicit in the scientific literature on the neural underpinnings of imagination. Let us consider the case of action. It has been a remarkable discovery that observing and executing an action involve (at least sometimes) the same resonance system in the brain, and more precisely the same “mirror neurons,” corresponding to types of action such as grasping, reaching, or eating (Rizzolatti et al. 1996; Rizzolatti et al. 2001). What about imagining an action? Marc Jeannerod claims that “imagining a movement relies on the same mechanisms as actually performing it, except for the fact that execution is blocked” (Jeannerod 2006, p. 28). Does this claim concern objective imagination, subjective imagination, or both? On the one hand, his notion of “motor imagery”, defined as “the ability to generate a conscious image of the acting self” (p. 23), strongly suggests that he is talking about subjective imagination. Motor imagery seems to underlie the imaginative recreation of an internal experience of action, such as the intimate experience we have while executing an action or controlling our bodily movements. On the other hand, Jeannerod makes clear that the “action representations” involved in motor imagery can also operate during action observation (p. 39). To the extent that visually imagining an action is analogous to observing an action, one may surmise that objective imagination too involves the relevant action representations.²³

What we would like to know, of course, is which action representations are common to both objective and subjective imagination of an action, and which action representations are specific to

²³ There is also the interesting case of observing one’s own action in a mirror. The question here is whether the observer is aware that she is observing her own action. If the answer is negative, then the re-creation of the relevant experience belongs to objective imagination. If the answer is positive, as for instances when one uses visual information to control one’s action (think of a man shaving in front of the mirror), then the re-creation of the relevant experience may also belong to subjective imagination (see footnote 11).

subjective imagination. Here as elsewhere, we think that phenomenological considerations can at least guide scientific investigations into the neural underpinnings of our ability to imagine actions, whether imaginatively observed or imaginatively executed.

7.3 Mindreading

We also think that much of the once-hot debate between the “theory theory” and the “simulation theory” of mindreading has missed the distinction between objective and subjective imagination, or at least its significance. Mindreading is often described as involving ways of “putting oneself in another person’s shoes” (Goldman 2006). However, as many have observed, that colloquial phrase can be used to refer to two different projects. One might try to understand either what one would do if one were in the other’s situation or what the other will do. The difference between these meanings has been conceived as depending on whether one performs the right “egocentric shift” and succeeds in mimicking the other’s mind (Gordon 1995). If we are right, there is another distinction that is crucial to simulation-based mindreading, namely the objective versus subjective imagination distinction. We might perform the right egocentric shift but imaginatively re-create only the other’s external experiences. For instance, we might imaginatively adopt the other’s visual point of view and try to understand what he or she is actually seeing. In doing so, though, we imaginatively adopt a perspective that is not necessarily the other’s perspective. Visual perspectives can be shared. It is only if we re-create at least some of the other’s internal experiences that we imaginatively adopt a perspective that can only be that of the agent. Unlike external experiences, internal experiences cannot be shared.

Why is it important for the success of mindreading that the mindreader re-creates also internal experiences of the other person? Let us consider the case of pain. To the extent that both objectively and subjectively imagining another person in pain may trigger the same resonance (affective) mechanisms, we can argue

that they are on par with respect to the imaginer’s *understanding* of the other’s experience (Gallese 2003). We surmise that the relevant difference between objectively and subjectively imagining the same painful experience concerns the *dynamics* of mindreading. Recreating an internal perspective on pain will spontaneously give rise to other subjective imaginings involving the recreation of the mental consequences of pain in the other. Objective imagination of another person in pain will likely develop in different directions. For instance, if we re-create a visual experience as of someone in pain, we will be inclined to re-create other visual experiences of the consequences of pain. More generally, someone who would be able to re-create only external experiences of pain would be blind to the internal consequences of pain. In contrast, subjective imagination promises to yield a better view of the other’s inner life as it unfolds in time.

7.4 Imaginative identification

In this essay, we did not explicitly mention an intriguing phenomenon in the field of imagination, namely our ability to *imagine being someone else*, or imaginative identification. For instance, we can imagine being Napoleon seeing the desolation at Austerlitz and being vaguely aware of one’s short stature (Williams 1976, p. 43). Recanati calls such cases “quasi-de se imaginings”:

I will, therefore, coin the term ‘quasi-de se’ to refer to the first person point of view type of thought one entertains when one imagines, say, being Napoleon. The type of imagining at stake is clearly first personal, yet the imaginer’s self is not involved [...]. The properties that are imaginatively represented are not ascribed to the subject who imagines them, but to the person whose point of view she espouses. (Recanati 2007, pp. 206–207)

How can an imagining be both first-personal and not genuinely (but only “quasi”) *de se*? If we can imagine being Napoleon just by recreat-

ing his visual experience of the desolation at Austerlitz, it is not obvious that quasi-*de se* imagination is necessarily first-personal. Since visual perspectives can be shared, our visual imagining can re-create anyone's perspective. In other words, objective imagination (i.e., the re-creation of external perspectives) would not be sufficient to generate quasi-*de se* imaginings. Perhaps Recanati implicitly ties quasi-*de se* imagination to subjective imagination so that imagining being someone else involves the recreation of at least some internal experience. Again, in contrast to external perspectives, internal perspectives cannot be shared. For instance, a subject imagining to be Napoleon might, on the one hand, see in imagination the desolation at Austerlitz (i.e., an external perspective is re-created) and, on the other hand, be vaguely aware of his short stature and his hand in his tunic (i.e., an internal, proprioceptive perspective is re-created).

In what sense would subjective imagination be first-personal, then? One view is that the quasi-*de se* case somehow derives from the genuine *de se* case, in which we imagine ourselves having various external and internal experiences. On this view, there is an asymmetrical dependence between quasi-*de se* and genuine *de se* imagination: even if the former is not merely a type of the latter, imagining being someone else having such-and-such experiences depends on the ability to imagine oneself having these experiences.

However, our account of subjective imagination suggests an alternative view, according to which the identity of the subject need not be built into a subjective imagining. Consider the case of action again. The constraint imposed on subjective imagination, that the imagined perspective on the action can only be that of the agent, leaves open whose self is involved. That the action is my action, or someone else's, is an additional fact in the imaginary world. In other words, subjective imagination can be neutral as to the identity of the self that occupies the relevant internal perspective. As a consequence, the same neutral imagining can give rise to either quasi-*de se* or genuine *de se* imagination, depending on

the imaginary project at stake.²⁴ Subjectively imagining oneself swimming and subjectively imagining another person swimming both rest on the same type of imagining, i.e., the recreation of an internal experience of the action of swimming. We take this neutrality to be a potential advantage for our analysis of subjective imagination. Subjective imagination can be seen as a basis for the introduction of a notion of self that is conceptually on a par with other selves. In this respect, imagination acts as an antidote to solipsism.

8 Conclusion

In this essay, we have tried to clarify what it means to claim that imagination is experiential. As we have seen, the notion of experiential imagination is not unitary and refers to a variety of phenomena. We have focused our attention on four aspects of this notion.

- First, experiential imagination broadly means that different kinds of experiential states are re-created in the imagination (although we have remained silent about the precise way in which the experiential states are re-created).
- Second, the distinction between external and internal experiences, which is independently motivated in the literature on non-imaginative mental states, has given rise to a helpful sub-division of experiential imagination into two different ways of imagining: objectively and subjectively. *Pace* Vendler, we have argued that this contrast cannot be straightforwardly aligned with two ways in which the self is involved in our imaginings (respectively, explicitly, or implicitly).
- Third, the literature commonly acknowledges two other varieties of imagination, namely sensory and cognitive imagination.

²⁴ The notion of imaginative project comes from Williams (1976). Imaginings are particular mental states, whereas imaginative projects can bind several imaginings in a coherent endeavour of imaginative world-making. The distinction is relevant even when a single imagining is at stake. Typically, an imaginative project will impose constraints, e.g., of an intentional or stipulative sort, on what is the case in the imagined world in addition to what is explicitly represented in an imagining.

We have pointed out that they should be considered as two sub-varieties of objective imagination, insofar as they both re-create external experiences (respectively, the five senses and at least some occurrent beliefs).

- Fourth, we suggested, more tentatively, that subjective imagination too may be further divided. There would be, on the one hand, the imaginative re-creation of *non-cognitive* non-sensory internal experiences (e.g., proprioception, agentive experiences, introspection, feeling pain) and, on the other hand, the imaginative re-creation of *cognitive* non-sensory internal experiences (e.g., ascent routines).

Of course, more has to be said about the precise domain of experiential states that can be re-created in the imagination, beyond those that we have introduced in this essay. Another question is whether there is something like non-experiential imagination. It might well be that, at the end of the journey, every type of imagining can be shown to belong to experiential imagination. This would have to include the state of imagining being a descendant of Napoleon, which, as we have seen, Walton rates as non-experiential. For instance, one might suggest that it is the state of imagining believing that one is a descendant of Napoleon (understood as representing in imagination a world in which one is a descendant of Napoleon).

Eventually, an analysis of experiential imagination on the lines suggested above should throw light not only on imagination *per se*, but on connected phenomena. As we have tried to illustrate, we believe that traditional and contemporary discussions about the relationship between imagination and possibility, the nature of mindreading, and the ability to imagine being someone else, often rely on oversimplified conceptions of imagination, and that a more fine-grained taxonomy of experiential imagination is needed. We suspect that our taxonomy is beneficial to still other applications of the notion of imagination, but we have to leave the task of justifying our suspicion to another occasion.

Acknowledgements

We thank the audience of the workshop “Self and Agency” in Liège (April 2014) for valuable feedback, two anonymous referees, and the editors Thomas Metzinger and Jennifer Windt for most helpful comments on earlier drafts of this essay.

References

- Anscombe, G. E. M. (1957). *Intention*. Oxford, UK: Blackwell.
- Arcangeli, M. (2011a). L'immaginazione ricreativa. *Sistemi intelligenti*, 23 (1), 59-74. [10.1422/34612](https://doi.org/10.1422/34612)
- (2011b). *The imaginative realm and supposition*. Paris, FR: University Paris 6-UPMC PhD Dissertation.
- Bayne, T. & Montague, M. (Eds.) (2011). *Cognitive phenomenology*. Oxford, UK: Oxford University Press.
- Bermúdez, J. L. (1998). *The paradox of self-consciousness*. Cambridge, MA: MIT Press.
- Bermúdez, J. L., Marcel, A. & Eilan, N. (1995). *The body and the self*. Cambridge, MA: MIT Press.
- Botvinick, M. & Cohen, J. (1998). Rubber hands “feel” touch that eyes see. *Nature*, 391 (6669), 756-756. [10.1038/35784](https://doi.org/10.1038/35784)
- Burge, T. (2005). Disjunctivism and perceptual psychology. *Philosophical Topics*, 33 (1), 1-78. [10.5840/philtopics20053311](https://doi.org/10.5840/philtopics20053311)
- Byrne, A. (2010). Recollection, perception, imagination. *Philosophical Studies*, 148 (1), 15-26. [10.1007/s11098-010-9508-1](https://doi.org/10.1007/s11098-010-9508-1)
- Carruthers, P. (2002). The roots of scientific reasoning: infancy, modularity, and the art of tracking. In P. Carruthers, S. Stich & M. Siegal (Eds.) *The cognitive basis of science* (pp. 73-95). Cambridge, UK: Cambridge University Press.
- Carruthers, P. & Veillet, B. (2011). The case against cognitive phenomenology. In T. Bayne & M. Montague (Eds.) *Cognitive phenomenology* (pp. 35-56). Oxford, UK: Oxford University Press.
- Casey, E. S. (1976). *Imagining: A phenomenological study*. Bloomington: Indiana University Press.
- Cassam, Q. (1999). *Self and world*. Oxford, UK: Oxford University Press.
- Chalmers, D. (2002). Does conceivability entail possibility? In T. S. Gendler & J. Hawthorne (Eds.) *Conceivability and possibility* (pp. 145-200). Oxford, UK: Oxford University Press.
- Cole, J. (1995). *Pride and a daily marathon*. Cambridge, MA: MIT Press.
- Crane, T. (2013). Unconscious belief and conscious thought. In U. Kriegel (Ed.) *Phenomenal intentionality* (pp. 156-173). Oxford, UK: Oxford University Press.
- Currie, G. & Ravenscroft, I. (2002). *Recreative minds: Imagination in philosophy and psychology*. Oxford, UK: Oxford University Press.
- De Vignemont, F. (2013). The mark of bodily ownership. *Analysis*, 73 (4), 643-651. [10.1093/analys/ant080](https://doi.org/10.1093/analys/ant080)
- Doggett, T. & Egan, A. (2007). Wanting things you don't want: The case for an imaginative analogue of desire. *Philosophers' Imprint*, 7 (9), 1-17.
- Dokic, J. (2008). Epistemic perspectives on imagination. *Revue Internationale de Philosophie*, 243 (1), 99-118.
- Dorsch, F. (2012). *The unity of imagining*. Berlin, GER: De Gruyter.
- Evans, G. (1982). *The varieties of reference*. Oxford, UK: Oxford University Press.
- Gallagher, S. (2005). *How the body shapes the mind*. Oxford, UK: Oxford University Press.
- Gallese, V. (2003). The manifold nature of interpersonal relations: The quest for a common mechanism. *Philosophical Transactions of the Royal Society of London*, 358 (1431), 517-528. [10.1098/rstb.2002.1234](https://doi.org/10.1098/rstb.2002.1234)
- Gendler, T. S. & Hawthorne, J. (Eds.) (2002). *Conceivability and possibility*. Oxford, UK: Clarendon Press.
- Gibson, J. J. (1966). *The senses considered as perceptual systems*. Boston, MA: Houghton Mifflin.
- Goldman, A. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. Oxford, UK: Oxford University Press.
- Gordon, R. (1995). Simulation without introspection or Inference from me to you. In M. Davies & T. Stone (Eds.) *Mental simulation* (pp. 53-67). Oxford, UK: Blackwell.
- Haggard, P. (2005). Conscious intention and motor cognition. *Trends in Cognitive Science*, 9 (6), 290-295. [10.1016/j.tics.2005.04.012](https://doi.org/10.1016/j.tics.2005.04.012)
- Husserl, E. (1901). *Logische Untersuchungen. Zweiter Teil: Untersuchungen zur Phänomenologie und Theorie der Erkenntnis*. Halle, GER: Max Niemeyer.
- Jeannerod, M. (2006). *Motor cognition. What actions tell the self*. Oxford, UK: Oxford University Press.
- Jeannerod, M. & Pacherie, E. (2004). Agency, simulation and self-identification. *Mind & Language*, 19 (2), 113-146. [10.1111/j.1468-0017.2004.00251.x](https://doi.org/10.1111/j.1468-0017.2004.00251.x)
- Kind, A. (2001). Putting the image back in imagination. *Philosophy and Phenomenological Research*, 62 (1), 85-109. [10.1111/j.1933-1592.2001.tb00042.x](https://doi.org/10.1111/j.1933-1592.2001.tb00042.x)
- (2011). The puzzle of imaginative desire. *Australasian Journal of Philosophy*, 89 (3), 1-19. [10.1080/00048402.2010.503763](https://doi.org/10.1080/00048402.2010.503763)
- Kosslyn, S. M. (1980). *Image and mind*. Cambridge, MA: Harvard University Press.
- (1994). *Image and brain: The resolution of the imagery debate*. Cambridge, MA: MIT Press.

- Kosslyn, S. M., Thompson, W. L. & Ganis, G. (2006). *The case for mental imagery*. Oxford, UK: Oxford University Press.
- Kripke, S. (1980). *Naming and necessity*. Cambridge, MA: Harvard University Press.
- Martin, M. (2002). The Transparency of Experience. *Mind and Language*, 17 (4), 376-425. [10.1111/1468-0017.00205](https://doi.org/10.1111/1468-0017.00205)
- McGinn, C. (2004). *Mindsight: image, dream, meaning*. Cambridge, MA: Harvard University Press.
- Meinong, A. (1902). *Über Annahmen*. Leipzig, GER: Verlag Johann Ambrosius Barth.
- Metzinger, T. (2003). *Being no one., The self-model theory of subjectivity*. Cambridge, MA: MIT Press.
- (2013). The myth of cognitive agency: subpersonal thinking as a cyclically recurring loss of mental autonomy. *Frontiers in Psychology*, 4 (931), 1-19. [10.3389/fpsyg.2013.00931](https://doi.org/10.3389/fpsyg.2013.00931)
- Mulligan, K. (1999). La varietà e l'unità dell'immaginazione. *Rivista di Estetica*, 11 (2), 53-67.
- Nichols, S. (2006). Imaginative blocks and impossibility: An essay in modal psychology. *The architecture of the imagination. New essays on pretense, possibility, and fiction* (pp. 237-256). Oxford, UK: Clarendon Press.
- Nichols, S. & Stich, S. (2003). *Mindreading: An integrated account of pretence, self-awareness, and understanding other minds*. Oxford, UK: Oxford University Press.
- Noordhof, P. (2002). Imagining objects and imagining experiences. *Mind and Language*, 17 (4), 426-455. [10.1111/1468-0017.00206](https://doi.org/10.1111/1468-0017.00206)
- O'Shaughnessy, B. (1980). *The will: a dual aspect theory*. Cambridge University Press: Cambridge, UK.
- Pacherie, E. (2007). The sense of control and the sense of agency. *Psyche*, 13 (1), 1-30.
- Peacocke, C. (1985). Imagination, possibility and experience: A berkeleian view defended. In J. Foster & H. Robinson (Eds.) *Essays on berkeley* (pp. 19-35). Oxford, UK: Oxford University Press.
- (2014). *The mirror of the world. Subjects, consciousness, and self-consciousness*. Oxford, UK: Oxford University Press.
- Perry, J. (2011). On knowing one's self. In S. Gallagher (Ed.) *The oxford handbook of the self* (pp. 370-391). Oxford, UK: Oxford University Press.
- Pylyshyn, Z. W. (2002). Mental imagery: In search of a theory. *Behavioral and Brain Sciences*, 25 (2), 157-182. [10.1017/S0140525X02000043](https://doi.org/10.1017/S0140525X02000043)
- Recanati, F. (2007). *Perspectival thought: a plea for (moderate) relativism*. Oxford, UK: Oxford University Press.
- (2012). *Mental files*. Oxford, UK: Oxford University Press.
- Rizzolatti, G., Fadiga, L., Matelli, M., Bettinardi, V., Paulesu, E., Perani, D. & Fazio, F. (1996). Localization of grasp representations in humans by PET. 1. *Experimental Brain Research*, 111 (2), 246-252. [10.1007/bf00227301](https://doi.org/10.1007/bf00227301)
- Rizzolatti, G., Fogassi, L. & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience*, 2 (9), 661-670. [10.1038/35090060](https://doi.org/10.1038/35090060)
- Schacter, D. L. & Addis, D. R. (2007). The cognitive neuroscience of constructive memory: remembering the past and imagining the future. *Philosophical Transactions of the Royal Society of London*, 362 (1481), 773-786. [10.1098/rstb.2007.2087](https://doi.org/10.1098/rstb.2007.2087)
- Tye, M. (1991). *The imagery debate*. Cambridge, MA: MIT Press.
- (2009). *Consciousness revisited: Materialism without phenomenal concepts*. Cambridge, MA: MIT Press.
- Vendler, Z. (1984). *The matter of minds*. Oxford, UK: Clarendon Press.
- Walton, K. L. (1990). *Mimesis as make-believe: on the foundations of the representational arts*. Cambridge, MA: Harvard University Press.
- Weinberg, J. & Meskin, A. (2006a). Imagine that! In M. Kieran (Ed.) *Contemporary debates in aesthetics and the philosophy of art* (pp. 222-235). Oxford, UK: Wiley-Blackwell.
- (2006b). Puzzling over the imagination: Philosophical problems, architectural solutions. In S. Nichols (Ed.) *The architecture of the imagination: New essays on pretence, possibility, and fiction* (pp. 175-202). Oxford, UK: Oxford University Press.
- White, A. (1990). *The language of imagination*. Oxford, UK: Blackwell.
- Williams, B. (1976). *Problems of the self: philosophical papers 1956-1972*. Cambridge, UK: Cambridge University Press.
- Williamson, T. (2008). *The philosophy of philosophy*. Oxford, UK: Blackwell.
- Windt, J. M. (2014). *Dreaming: A conceptual framework for philosophy of mind and empirical research*. Cambridge, MA: MIT Press.
- Wollheim, R. (1984). *The thread of life*. Cambridge, UK: Cambridge University Press.

Imagination and Experience

A Commentary on Jérôme Dokic & Margherita Arcangeli

Anne-Sophie Brüggén

Jérôme Dokic and Margherita Arcangeli develop a taxonomy of the mental states classified as experience-like imaginings in their paper “The Heterogeneity of Experiential Imagination”. Experience-like imaginings are thought to *re-create* experiences. Therefore, the taxonomy of the Experiential Imagination suggested by the authors mirrors a taxonomy of the underlying, re-created experiences. In this commentary, I will focus on the notion of re-creation that is invoked, and argue that this notion must either be fleshed out further or omitted from the taxonomy. Two further points follow this discussion: first I will discuss the idea of different kinds of self-involvement in objective and subjective imagination and suggest an alternative view. Then I raise some doubts about the classification of cognitive imaginings as experiential imaginings. To summarise, I will suggest an alternative interpretation of these findings by claiming that we can obtain a useful taxonomy of imaginative states based on our pre-theoretical opinions. Furthermore, I will explore the idea that experiential imaginings involve an empty point of view.

Keywords

Cognitive imagination | Experiential imagination | Objective imagination | Sensory imagination | Subjective imagination

Commentator

Anne-Sophie Brüggén

anne-sophie.brueggen@gmx.net

Target Authors

Jérôme Dokic

dokic@ehess.fr

Institute Jean Nicod
Paris, France

Margherita Arcangeli

margheritarcangeli@gmail.com

Institute Jean Nicod
Paris, France

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

In their paper “The Heterogeneity of Experiential Imagination”, Jérôme Dokic and Margherita Arcangeli offer a taxonomy of the various mental states subsumed by them under the label Experiential Imagination. Experiential Imagination is introduced as the re-creation of non-imaginative, conscious mental states. Since experiential imaginings re-create experiential mental states, they can be classified according to the underlying taxonomy of the conscious mental states that they re-create. Dokic and Arcangeli

argue that there are two types of Experiential Imagination: objective imagination and subjective imagination. Objective imagination re-creates experiences about the external world, while subjective imagination re-creates experiences about mental or bodily states of oneself. Furthermore, the authors refine the category of the objective imagination by dividing it into sensory imagination and cognitive imagination. This taxonomy of the Experiential Imagination suggested by Dokic and Arcangeli provides a struc-

ture within which to understand the vast spectrum of mental states classified as experiential imaginings by referring to the notions of subjective and objective imagination. The authors additionally suggest an attractive perspective on cognitive imaginings, which relies on the idea that these have a phenomenal character as well.

I would like to discuss three aspects of Dokic and Arcangeli's paper and close with my own reflections on the topic. I will start with a point concerning the definition of Experiential Imagination as re-creating other mental states (section 2). Two points about the taxonomy itself will follow this discussion: the second point deals with the notions of objective and subjective imagination (section 3). A third point with which I will be concerned is the classification of cognitive imaginings within the suggested taxonomy (section 4). It is unclear whether and in what sense the notion of *re-creation* is helpful for delineating the suggested taxonomy of Experiential Imagination. The taxonomy faces certain issues that are partly grounded in the notion of *re-creation*.

Given these considerations, I will present my own take on a classification of imaginings that does not involve the notion of re-creation and is based on our pre-theoretical opinions about imaginings. In addition to this, I explore the notion of an *empty perspective* to describe a phenomenological difference in the perspectival character of imaginings and non-imaginative experiences (see section 5).

2 Re-creating experiences in imagination

I would like to focus first on the notion of Experiential Imagination itself. Dokic and Arcangeli want to develop a taxonomy of Experiential Imagination, and they therefore start by exploring the mental states that fall under this category. The authors introduce the subject of their taxonomy, the Experiential Imagination, as follows (see Dokic & Arcangeli [this collection](#), p. 2): Experiential Imagination is first of all imagination that is *experience-like*. Whether all instances of imaginings are of this kind or whether there may be kinds of imagination that do not fall under this category is left open

(Dokic & Arcangeli [this collection](#), p. 2). The notion of Experiential Imagination is spelled out further by referring to Christopher Peacocke's so-called General Hypothesis (GH):

To imagine something is always at least to imagine, from the inside, being in some conscious state (see Peacocke 1985, p. 21).

According to this definition, Experiential Imagination is imagining something *from the inside*, which is defined as involving "the perspective of a conscious experience" (Dokic & Arcangeli [this collection](#), p. 3). An example would be visually imagining a white sandy beach, which involves a certain experiential perspective (Dokic & Arcangeli [this collection](#), p. 3). The authors call this kind of imagination "X-like" imagination or "re-creating X" in imagination (Dokic & Arcangeli [this collection](#), p. 3), with X standing for the non-imaginative mental state that is re-created (Dokic & Arcangeli [this collection](#), p. 3). Following this terminology, visually imagining a white sandy beach is vision-like imagination or re-creating a visual experience of a white sandy beach in the imagination. The authors sum up these considerations in a brief discussion on the notion of *re-creation*: Experiential Imagination is, according to the authors, imagination that re-creates non-imaginative conscious states (Dokic & Arcangeli [this collection](#), p. 3). The idea that imaginative states re-create other mental states allows Dokic & Arcangeli to ground their taxonomy of the Experiential Imagination on a classification of such re-created mental states. A taxonomy of these underlying non-imaginative mental states can therefore serve as a basis for a taxonomy of the corresponding imaginative states ([this collection](#), p. 3). Dokic and Arcangeli do not commit themselves to any existing account that explains the imagination in terms of *re-creation* or *simulation* (Dokic & Arcangeli [this collection](#), p. 3). The notion of re-creating a non-imaginative mental state is not explored further, since "it is enough for our purposes to accept the idea that a phenomenologically useful taxonomy of imagination can be guided

by a corresponding taxonomy of non-imaginative mental states” (Dokic & Arcangeli [this collection](#), p. 3).

Even if the authors wish to remain as neutral as possible with respect to the notion of re-creation, it is important to spell it out. There are two main reasons why I think that this notion should be explored further: first, the notion of re-creation is crucial to the nature and scope of the taxonomy in which it is involved. Second, it seems to me that the authors oscillate to some extent between different notions of re-creation, rather than actually remaining neutral about it.

Concerning the first point, there seem at least three options available for understanding the idea that imaginings re-create other mental states, assuming that *re-creating* is not used to specify sub-personal processes but deals instead with mental states on a personal level:

- (1) As a *mere way of speaking* to refer to x-like imaginings
- (2) As the claim that imaginings re-create an experiential *mode*
- (3) As the claim that imaginings re-create experiences as part of their *contents*

The first way to understand the notion of re-creating is to use it synonymously with the notion of x-like imagination. What I mean by this is that we may use the notion of *re-creating X in imagination* to refer to *having an imagining with an x-like phenomenology*. In which case, for example, *re-creating a visual experience in imagination* would be synonymous with *having a vision-like phenomenology*. Understood like this, the notion of re-creating is simply used to refer to imaginings with an experience-like phenomenology. This is merely a way of speaking or a terminological stipulation. If the notion is used like this, it does not assume or specify any relation between imagination and experience in general (or between particular imaginings and experiences). That is, using the notion in this way does not commit us to the claim that imaginings are related to or dependent on experi-

ences in any sense. However, if the notion of re-creation is used as a mere way of speaking, it would be better to omit it from the taxonomy altogether, since it does not play any explanatory role or add any technical term. Instead, we could simply speak of *x-like* imagination and thereby refer to imaginings that have an x-like phenomenology.

The other two ways of spelling out the notion of re-creating are more substantial than just synonyms for x-like imaginings: in these versions, the notion of re-creation is a *metaphysical notion* that is used to indicate a relation between imaginings and experiences. Used like this, the notion of re-creation involves a claim about the metaphysical structure of imaginings (or the imagination), since it endorses the idea that imaginings are related to experiences in a specific way. The nature of this relation can be spelled out differently. Version (2) claims that imaginings re-create experiences in the following sense: for every type of experience there is a respective *imaginative mode*. There is a visual mode of imagination, an auditory mode of imagination, a proprioceptive mode of imagination, and so forth. In this sense, every type of experience is re-created by a specific type of Experiential Imagination. Version (3) claims something else, namely that different experiences are re-created as part of the contents of imaginings: if I visually imagine an object O, for example, the imagining has as part of its content a visual experience of O.

These two notions of re-creation yield different taxonomies with different metaphysical underpinnings: a taxonomy based on (2) differentiates imaginings according to their *mode*, while a taxonomy based on (3) classifies imaginings according to their *contents*. If re-creation is understood as specified in (2), such that for every experience-type there is an imaginative type that re-creates this experience-type, this is a different metaphysical claim to the one sketched in (3). As such, one could claim that there is one type of imagination that re-creates various experience-types by taking them up as parts of their contents. The nature of the relation called *re-creation* therefore has consequences for what is taxonomised: this can be,

for example, the mode or the content of an imagining. Neglecting this notion (if it is considered to be a substantial metaphysical notion) therefore means neglecting the metaphysical basis of the taxonomy. Thus, it seems to me that from a methodological point of view it is indeed important to clarify which notion of re-creation is in play.

The second worry I want to raise about the notion of re-creation is that the authors do not in fact remain neutral with regard to this notion. First, it seems that the notion of re-creating that the authors have in mind is not only a synonym for the expression *x-like imaginings*. One reason to think so is that Dokic and Arcangeli use the notion of re-creation in crucial definitions such as, for example, to formulate the various versions of the General Hypothesis. One example is as follows:

SensH: To imagine something sensorily is always at least to re-create some sensory experience. (Dokic & Arcangeli [this collection](#), p. 4)

If *to re-create some sensory experience* is synonymous with *having an imagining with a sensory phenomenology*, the hypothesis and its variants are no longer interesting claims. This indicates that the notion is more than what I called a mere way of speaking, but instead refers to (and thereby stipulates) a relation between imaginings and experiences or imagination and experience in general.

Additionally, it seems to me that the suggested taxonomy oscillates between different notions of re-creation. On the one hand, Dokic and Arcangeli sometimes seem to sympathise with the mode-sense of the notion of re-creation (as in (2)). When introducing the distinction between objective and subjective imagination, they claim, for example, that this distinction is concerned with the mode of the experience and not with the content (Dokic & Arcangeli [this collection](#), p. 9). I address this point in more detail in section 3, below. On the other hand, Dokic and Arcangeli employ the General Hypothesis and develop various variants of it. As a reminder, the General Hypothesis claims that

“to imagine something is always at least to imagine, from the inside, being in some conscious state” (Peacocke 1985, p. 21). This thesis is put forward by Christopher Peacocke (1985, p. 21) and Michael Martin (2002), who call it the “Dependency Thesis” (Martin 2002). It is usually considered to be a claim about what an imagining represents (see e.g., Dorsch 2012, pp. 294 and pp. 314; see also Paul Noordhof’s exploration and criticism of the thesis in Noordhof 2002). The idea behind these claims is that imaginings are experiential in nature because what we imagine in the imagining are experiences: “sensory imagining is experiential or phenomenal precisely because what is imagined is experiential or phenomenal” (Martin 2002: 406). This means that my visual imagining of an object *O* *represents* an experience of *O* and therefore is experiential. The General Hypothesis hence seems to imply, at least implicitly, a specific conception of re-creation: it endorses the idea that imaginings involve experiences as part of their contents, which is the notion of re-creation I formulated in version (3). Therefore, this view is not neutral about the nature of re-creating: relying on the General Hypothesis brings with it a certain commitment about the notion of re-creation involved (given that one adopts the suggested reading of the General Hypothesis and its variants).

In this section, I (1.) discussed three interpretations of the notion of re-creation that I take to be the most relevant in the given context, since they are alluded to by the authors. It seems that the notion of re-creation needs to be fleshed out further if it is to play some explanatory role in the taxonomy (otherwise it can be dismissed); and (2.) argued that the background assumptions of the taxonomy are committed to differing interpretations of the notion of re-creation. Therefore, the authors do not remain neutral about the notion of re-creation that is involved here but seem to implicitly adopt different notions of re-creation. One way of solving these issues would be to address them and commit to a specific notion of re-creation. Another solution would be to eliminate the notion of re-creation from the taxonomy, which is what I will suggest in the final section of this commentary.

3 Subjective and objective imagination and the self

One central aspect of the taxonomy that Dokic & Arcangeli propose is the distinction between subjective imagination and objective imagination (see [this collection](#), pp. 4). Subjective imagination re-creates internal experiences: experiences that are “supposed to be about a mental or bodily state of oneself” (Dokic & Arcangeli [this collection](#), p. 6). As an example, the authors point to “proprioceptive and agentic experiences” (Dokic & Arcangeli [this collection](#), p. 6) such as imagining the movements of swimming in the sea. In contrast, objective imagination re-creates external experiences. These are experiences that are “typically about the external world” (Dokic & Arcangeli [this collection](#), p. 6)—such as, for example, visual experiences of objects. Dokic & Arcangeli claim that experiential imaginings in general can be divided into subjective and objective imaginings ([this collection](#), p. 6). In a second step, this differentiation is then distinguished from Zeno Vendler’s distinction between imaginings that either implicitly or explicitly involve the self (Dokic & Arcangeli [this collection](#), pp. 7). The authors argue that Vendler’s categorisation differs from their own by providing four examples of cases of subjective and objective imagination that involve the self either implicitly or explicitly (Dokic & Arcangeli [this collection](#), p. 8).

I have a number of worries about some of the ideas and notions that the authors put forward along this line of thought. My first worry concerns the claim that the suggested differentiation of objective and subjective imagination concerns the *mode* of the respective state and therefore differs from Vendler’s distinction, which is thought to be about the state’s *content* (Dokic & Arcangeli [this collection](#), p. 8). Internal and external experiences are equally *internal* in some sense, since they are experiences that are *internal to some subject*. As I understand the authors here, the difference between internal and external experiences is that they are usually about internal or external entities, respectively. Thus, in the given context, the notions *internal* and *external* apparently specify

what the experiences are about. On the level of imagination, subjective and objective imagination re-creates these different types of experiences. The authors specify this idea by spelling out two versions of the General Hypothesis adapted for objective and subjective imagination, called *ObjH* and *SubjH* (Dokic & Arcangeli [this collection](#), p. 6). As I specified above in section 2, one can read the General Hypothesis and its variants as claiming that imaginings re-create experiences in the sense that they represent experiences as part of their contents. If one accepts this interpretation, it is not obvious to me why and how re-creating internal and external experiences in the imagination yields imaginings that are different in *mode* (namely subjective and objective imaginings) and not in terms of what they represent. This point is an exemplification of the issue I raised in section 2: it depends on how one spells out the notion of re-creation whether or not the line of argument that the authors present to distinguish their notions from Vendler’s is convincing.

My second worry concerns the notion of *implicitly involving the self*. It seems to me that there is room to argue that both objective and subjective imagination as defined by Dokic and Arcangeli always involve the self implicitly (the authors briefly address this point in footnote 13). If this were the case it is unclear how their notions are different from Vendler’s. The self is implicitly involved in an imagining if “it fixes the point of view internal to the imagined scene without being a constituent of that scene” (Dokic & Arcangeli [this collection](#), p. 7). An example is imagining seeing the Pantheon: there is a specific point of view involved in this imagining (Dokic & Arcangeli [this collection](#), p. 7). This, however, seems to be the definition of Experiential Imagination *in general* that the authors propose in the beginning of the paper. They explain (by referring to Peacocke) that Experiential Imagination always involves an experiential perspective (Dokic & Arcangeli [this collection](#), p. 3). If involving an experiential perspective is sufficient to implicitly involve the self, and if experiential imaginings are defined as imaginings that involve an experiential perspective, then every experiential imagining in-

volves the self implicitly. If this is indeed how the authors conceive of Experiential Imagination, a notion introduced by Michael Martin may be helpful for dismissing certain difficulties (though I am aware that he uses this notion in a context with different argumentative aims). [Martin](#) argues (similarly to Peacocke) that at least some sensory imaginings involve a point of view, and thereby implicitly represent experiences (2002, pp. 40). However, as he explains, the presence of a point of view in the imagining does not imply that I myself occupy this point of view: “[t]he point of view within the imagined scene is notoriously empty enough that one can in occupying that point of view imagine being someone other than one actually is” ([Martin 2002](#), p. 411). I take this to be a promising way of differentiating imaginings from non-imaginative experiences, since they involve different kinds of points of view or perspectivalness (I will say more on this in section 5).

Maybe this notion of an *empty* point of view can also be helpful for further sharpening the notions of objective and subjective imagination. One could argue that objective experiential imaginings involve a point of view—but an empty one. Thus, imagining seeing the Pantheon involves a point of view, but this point of view is empty in the sense that it must not be myself occupying this point of view. In this sense, objective imaginings may not involve the self at all. This observation could also serve to set the subjective/objective distinction apart from Vendler’s. But it is probably more difficult to transfer the notion of an empty point of view to subjective imagination, given that it is defined as re-creating experiences about oneself. Maybe this is close to what the authors have in mind when they loosen the notion of subjective imagination towards the end of the paper by claiming that subjective imaginings may be neutral about the identity of the self involved ([Dokic & Arcangeli this collection](#), p. 16). Thus, to conclude, considering the notion of an empty point of view at least seems to be an interesting option to be explored in order to strengthen the objective/subjective distinction and the notion of subjective imagination. Apart from this suggestion, I will come back to the notion of an

empty point of view in the final section of this commentary and on this basis offer an additional perspective.

4 The phenomenal character of cognitive imaginings

My third and final point concerns the classification of cognitive imaginings. Cognitive imaginings are usually considered to be non-sensory in the sense of not having a sensory phenomenal character or indeed any phenomenal character at all. An example of cognitive imagination is to imagine that there is a largest prime number. [Dokic](#) and [Arcangeli](#) suggest that this orthodox classification may be misguided, since one can plausibly argue that cognitive imaginings have a certain phenomenology, namely a cognitive one ([this collection](#), pp. 10–11). Therefore, the authors claim, we could classify them as experiential imaginings as well.

I think the idea of ascribing a certain cognitive phenomenology to cognitive imaginings is very attractive, since it acknowledges the idea of a cognitive phenomenology in general and allows us to classify all kinds of imaginings according to one single feature, which is their phenomenal character (see also section 5). However, I am unsure about the classification of cognitive imaginings as experiential imaginings. Here is why: in the beginning of the paper, the authors define one important feature of the kinds of imaginings that they consider experiential: they involve an “experiential perspective” and are (in this sense) “from the inside” (see [Dokic & Arcangeli this collection](#), p. 3). It is not spelled out in detail how we should understand the notion of an experiential perspective but, as I interpret it, this involves at least that things are oriented “within egocentric space” ([Martin 2002](#), p. 408), to use Martin’s expression. Martin only speaks about visual perceptual experiences, but it seems to me that one can plausibly expand this notion to all kinds of experiences: they involve an egocentric perspective. As I understand Dokic and Arcangeli, they consider this egocentric perspective to be a defining feature of the phenomenology of experiential imaginings that re-create experiences.

If cognitive imaginings are considered to be experiential imaginings, and if experiential imaginings are considered to involve an egocentric perspective, one would expect cognitive imaginings to also have this egocentric perspective. However, it seems to me that the phenomenal character of cognitive imaginings does not involve the *perspective of an experience*. If I imagine that the earth is flat (and according to the authors thereby re-create the belief that the earth is flat) it seems that imagining this does not involve any egocentric perspective in the sense given above. If at all, cognitive imaginings incorporate a very specific kind of perspective that is distinct from any experiential *perspective*. Consequently, even if cognitive imaginings have a phenomenal character, this seems quite different from the phenomenal character of experiences (given that the latter is considered to involve an experiential perspective). If the authors endorse a different notion of experiential phenomenal character and *having an experiential phenomenal character* is, for example, just a synonym for *having a phenomenal character*, then my point is not valid. However, if Dokic and Arcangeli indeed think that having an experiential phenomenal character means that an egocentric perspective is involved (as in the case of experiences), I suggest that we need to reconsider the classification of cognitive imaginings as provided here. While I find the idea that cognitive imaginings may have some kind of phenomenal character convincing, it seems less convincing to me that they have an experiential phenomenal character in the sense discussed here. Therefore, I propose that we instead classify cognitive imaginings as a different kind of imagination with a specific cognitive phenomenal character.

5 Conclusion

The issues I raised in the previous sections can probably all be met in order to maintain the taxonomy suggested by Dokic and Arcangeli and to develop it further. Nevertheless, I think that the points I raised also allow for an alternative interpretation that offers a different perspective on a taxonomy of imaginings. Before

summarising the results of this commentary, I would like to explore this alternative perspective on the topic. My two main claims are: (1.) that it is not helpful to involve the notion of recreation in a taxonomy of imaginings, and that the taxonomy can be yielded without it; and (2.) that the specific way the self is (not) involved in imaginings *distinguishes* them from experiences rather than mirroring experiences.

Concerning the first point, it is neither necessary nor helpful to involve the notion of recreation or any other metaphysical notion if the aim is to yield a *phenomenological taxonomy* of imaginative states (and I take this to be one of the aims of Dokic and Arcangeli's paper). In order to yield such a phenomenological taxonomy, we can simply rely on our pre-theoretical classifications of imaginings as vision-like or action-like, and so forth. The notion *vision-like* and its cognates *x-like* can be understood as phenomenological notions here: to the imagining subject, what it is like to visually imagine an object is similar to what it is like to visually experience an object. That there are such similarities in phenomenal character is an interesting observation that allows us to build a phenomenological taxonomy. If one additionally accepts the idea of a cognitive phenomenology, this account allows us to capture cognitive imaginings as well, and to classify them according to their (cognitive) phenomenal character. Explaining *why* imaginings are vision-like or action-like, and what the metaphysical underpinnings of this phenomenological taxonomy may be is another task. These tasks should not be entangled.

One may worry that these pre-theoretical notions (such as *vision-like*) and opinions are too imprecise and not apt to yield a taxonomy of imaginative states that can ground further philosophical theorising. One answer to this worry is to expand a line of thought suggested by Fabian Dorsch. He considers the fact that we stably, effortlessly, and consistently “do group together a large variety of mental occurrences in the class of imaginings, while excluding many others” (Dorsch 2012, p. 6) to justify the idea that imaginings form a unified class of mental states. This line of thought can be adapted to ground a more fine-grained taxonomy of ima-

imaginings, based on our pre-theoretical opinions: we also stably, effortlessly, and consistently classify various imaginings as vision-like, audition-like, movement-like, and so forth. There are certainly borderline cases or instances of imaginings that combine several phenomenological aspects, but nevertheless this pre-theoretical classification is stable in the way described by Dorsch. I consider therefore this intuitive and pre-theoretical classification a helpful taxonomy of imaginings that can serve as a sufficiently justified *starting point* for further philosophical reflection. This pre-theoretical classification of imaginings that I suggest probably does not yield essentially different categories to the taxonomy suggested by Dokic and Arcangeli. It classifies imaginings according to their phenomenal character as vision-like, action-like, and so forth, which are all categories acknowledged by the authors. What I wish to claim is that in order to ground this taxonomy, it is not necessary or helpful to involve a metaphysical notion such as re-creation. It is sufficient to recur to our pre-theoretical classification of imaginative states.

The only category that is probably not reflected in this phenomenological taxonomy is the distinction between subjective and objective imagination, which, according to the authors, also “gives rise to phenomenologically different imaginings” (Dokic & Arcangeli [this collection](#), p. 6). The reason for this is that there is a difference between the more fine-grained phenomenology and the more coarse-grained phenomenology of a mental state. By this I mean that we can distinguish various aspects of a mental state’s phenomenal character. Two different visual experiences of a red apple and a green apple respectively share the coarse-grained phenomenal character of being visual, but they differ in terms of their fine-grained phenomenal character: perceiving a red apple is phenomenally different from perceiving a green apple. The taxonomy I suggest above is concerned with the rather coarse-grained phenomenal character of imaginings that allows us to classify them as vision-like, action-like, and so forth. An even more coarse-grained phenomenal character would be the one which all types of

imaginings have in contrast to cognitive state, for example. The distinction between objective and subjective imagination seems to reflect more fine-grained phenomenological categories than those that classify imaginings according to what their phenomenal character resembles. I am not sure whether there is a *phenomenology of objectiveness* (as opposed to subjectiveness) that, for example, unifies sensory imagination and cognitive imagination as opposed to proprioceptive imagination (as suggested by Dokic and Arcangeli). This shows that the account and methodology that I propose also faces certain challenges. One challenge would be to single out exactly which aspects of the phenomenology we take to be defining marks for a categorisation. Another challenge, for example, would be to point out that for this account we have to rely on introspective findings, whose epistemic status and reliability may be controversial. Nevertheless I think that pre-theoretical reflection based on phenomenological findings is an appropriate way to lay out a taxonomy of the mental states we classify as imaginings, since in principle it can be done stably, effortlessly, and consistently (see again [Dorsch 2012](#), p. 6).

The second aspect I would like to address is the distinction between subjective and objective imagination. These notions introduced by Dokic and Arcangeli are very helpful, since they reveal the particular ways in which the self (or aspects of the self) is involved in imaginings. However, I think one can draw different conclusions from these observations than those presented by the authors. As I suggested in [section 3](#), I think the best way to describe the point of view involved in imaginings is by adopting and expanding the notion of an *empty point of view*. It seems to me that imaginings do not involve the self in the same way as, for example, experiences do. I will explore this line of thought by pointing to the example of visual experiences as opposed to visual imaginings. The perspectival character of a visual experience has several aspects: it involves a distinct point of view that locates the perceiving subject in a determinate relation to its surrounding objects. Objects are therefore perceived as being close, far away, to

the left, above, and so forth (see also [Martin 2002](#), p. 408). In this sense the self is involved, since there is always an egocentric perspective. However, in imagination this kind of perspectivalness need not be fully realised. It seems possible to imagine an object without imagining it at a certain distance or at a certain position. If I perceive a tree, I perceive it far away to the left, for example. If I imagine a tree I can simply imagine the tree. I *can* imagine a tree in the distance to the left but this is something I deliberately add to the imagining. This thought can be expanded to other forms of imaginings as well. One way to capture this particular perspectival character of imaginings is to adopt the proposed notion of an empty point of view: while experiences involve the self in the sense of involving an egocentric perspective (which is a non-empty point of view), imaginings involve an *empty point of view*. This does not mean that one adopts, in imagining, the point of view of someone else (as opposed to the point of view of myself), but that this point of view is *empty*. One important difference between this notion of an empty point of view and Dokic and Arcangeli's account is that it *differentiates* imaginings from experiences: regarding the point of view that is involved, imaginings differ importantly from non-imaginative experiential states, since the former may involve an empty point of view. In contrast to this, Dokic and Arcangeli seem to think that imaginings mirror non-imaginative states with respect to the nature of the point of view involved (again probably partly due to the notion of re-creation). Again, the approach that I suggest certainly faces challenges. One challenge is to demand that we spell out the notion of an empty point of view in more detail. So far, I have only pointed in the direction of how to capture certain particular features of imaginings. However, investigating this difference further seems like a promising way to clarify the nature of imaginings.

To sum up, I will briefly repeat the points I discussed in this commentary:

1. I suggested that we explore the notion of re-creation further, since it occupies a central place in the suggested taxonomy of Experien-

tial Imagination. As I argued, this notion must either be spelled out or omitted from the taxonomy, since as an underdetermined notion it does not add to the explanatory basis. Furthermore, I showed that the authors seem to implicitly rely on different notions of re-creation instead of remaining neutral about it.

2. I pointed to some worries about the distinction between subjective and objective imagination. I suggested that we adopt the notion of an empty point of view to characterise the kind of self-involvement we find in experiential imaginings.
3. I formulated my doubts about the classification of cognitive imaginings as experiential imaginings due to their phenomenal character, which does not seem to be experiential in the sense that it does not involve an experiential perspective.

I concluded these considerations with my own interpretation of the findings. As I suggested, we can develop a phenomenological taxonomy of different types of imaginings by basing it on our pre-theoretical opinions about imaginings. We do not need to involve the notion of re-creation (or other non-phenomenological notions) in order to do this. Clarifying the metaphysical underpinnings of this taxonomy is a different task. Additionally, I interpreted reflections on the various ways the self is involved in imaginings as yielding the conclusion that imaginings differ from experiences in terms of how the self is (not) involved, rather than mirroring experiences, in this respect. Imaginings involve an empty point of view, while experiences have an egocentric point of view. I consider both these aspects relevant for any theory of imaginings.

Dokic and Arcangeli's taxonomy has essentially contributed to further developing a theory of imaginings by revealing and illuminating relevant aspects of the nature of imaginings. Their observations have clearly uncovered a neuralgic aspect of imaginings, which is how the self is involved (or not involved) in imaginings. Furthermore, their taxonomy allows us to classify cognitive imaginings in terms of their phenomenal character and not, for example, with respect to

what these are about. Although the taxonomy reveals how heterogeneous imaginings are, it therefore nevertheless offers a unified take on imaginings. Adopting Dokic and Arcangeli's observations as a starting point for further investigations will certainly be very fruitful, and is sure to advance our understanding of the nature of imaginings.

Acknowledgements

I want to thank the editors of this volume and the anonymous reviewers for their valuable comments, which have all been very helpful.

References

- Dokic, J. & Arcangeli, M. (2015). The heterogeneity of experiential imagination. In Metzinger, T. and Windt, J. M. (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Dorsch, F. (2012). *The unity of imagining*. Frankfurt a. M., GER: Ontos.
- Martin, M. G.F. (2002). The transparency of experience. *Mind and Language*, 17 (4), 376-425.
[10.1111/1468-0017.00205](https://doi.org/10.1111/1468-0017.00205)
- Noordhof, P. (2002). Imagining objects and imagining experiences. *Mind and Language*, 17 (4), 426-455.
[10.1111/1468-0017.00206](https://doi.org/10.1111/1468-0017.00206)
- Peacocke, C. (1985). Imagination, experience and possibility: a Berkeleian view defended. In Foster J. and Robinson H. (Eds.) *Essays on Berkeley* (pp. 19-35). Oxford, UK: Clarendon Press.

The Importance of Being Neutral: More on the Phenomenology and Metaphysics of Imagination

A Reply to Anne-Sophie Brügger

Jérôme Dokic & Margherita Arcangeli

In this reply to Anne-Sophie Brügger's comments to our target paper, we focus on three main issues. First, we explain that although our account of imaginative re-creation is in many respects metaphysically neutral, it allows for a taxonomy of imaginings that goes beyond mere phenomenological observations and pre-theoretical intuitions. Second, we defend our interpretation of the distinction between objective and subjective imagination and compare it with Brügger's own suggestions involving the notion of an empty point of view. Third, we insist that the notion of experiential perspective should be construed broadly and include cognitive or belief-like imagination.

Keywords

Cognitive imagination | Empty point of view | Objective imagination | Phenomenology | Re-creation | Subjective imagination

We would like to thank Anne-Sophie Brügger for her very interesting comments on our paper. In what follows, we try to respond to what we see as the central points raised in her discussion.

1 On the notion of re-creation

In our target paper, we use a notion of re-creation in order to set up a sophisticated taxonomy of experiential imagination. We also profess a certain neutrality with respect to this notion. Anne-Sophie

Brügger argues that our neutrality is only apparent, and that we in fact oscillate between two substantial notions of re-creation, which have quite different implications for the ontology of imaginings.

Authors

Jérôme Dokic

dokic@ehess.fr

Institute Jean Nicod
Paris, France

Margherita Arcangeli

argheritarcangeli@gmail.com
Institute Jean Nicod, France

Commentator

Anne-Sophie Brügger

anne-sophie.bruegger @ gmx.net

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu
Monash University
Melbourne, Australia

Our professed neutrality concerns only the subpersonal underpinnings of imagination. We do not want to commit ourselves to the view that imaginings and their non-imaginative counterparts share neural or functional resources. We do not explicitly vindicate any neutrality with respect to the notion of re-creation at the personal level. However, we intend to be neutral at that level too, in the following respect. In our account, the phrase “X re-creates Y” should be used synonymously with the phrase “X is Y-like”, to mean that an imagining of type X has a phenomenal character analogous to the phenomenal character of a non-imaginative state of type Y. For instance, visual imagination is visual-like in the sense that its phenomenal character is more similar to visual perception than, say, auditory perception or belief. In general, what matters for our purposes is that there is a systematic correspondence between the imaginative and the non-imaginative realms; the metaphysical nature of this correspondence is left open.

Now, Brüggen raises an interesting question, namely whether (notwithstanding our intentions) our account shows an oscillation between two different metaphysical conceptions of re-creation. On the first (mode-based) conception, there are different imaginative *modes* corresponding to kinds of experience in the non-imaginative realm. On the second (content-based) conception, which Brüggen attributes to Mike Martin, all imaginings belong to a single imaginative mode but represent different kinds of experience as part of their *contents*.

Brüggen suggests (following Martin’s 2002 interpretation) that Peacocke’s General Hypothesis (1985) already carries a commitment to the content-based conception. We disagree. The phrase “imagining being in some conscious state” (Peacocke 1985, p. 21) does not obviously entail that the conscious state is represented in the content of the imagining. It is compatible with taking the expression “being in some conscious state” to be a modifier of “imagining”, just as the internal accusative “a song” is a modifier of “singing” in “singing a song”. Perhaps we are wrong about Peacocke’s intentions, but we insist that our use of the General Hypo-

thesis can be metaphysically neutral in this sense.

What about the mode-based conception of re-creation? We concede that some of our formulations, especially when we introduce the distinction between objective and subjective imagination, evoke such a conception. As it happens, we have both rejected the content-based conception in other works (Dokic 2008; Arcangeli 2011a, 2011b). However, many aspects of our taxonomy can be re-formulated in terms more amenable to the latter conception. For instance, the distinction between objective and subjective imagination might be construed as a distinction between imaginings that represent external experiences and imaginings that represent internal experiences as part of their contents. Whether all aspects of our taxonomy can be re-formulated in this way is indeed something that should be explored further.

Brüggen eventually recommends getting rid of the notion of re-creation, and going for a purely phenomenological taxonomy based on pre-theoretical intuitions. It is worth contrasting our methodology with hers. In many respects, our taxonomy rests on well-identified phenomenological types. For instance, all visual imaginings are clearly unified under a single phenomenological type. The latter can then easily be related to a kind of experience in the non-imaginative realm, namely visual experiences. In other cases, identifying non-imaginative counterparts is more difficult because the relevant imaginings do not form a well-identified phenomenological type. We agree with Brüggen that there may not be a phenomenology of objective (as opposed to subjective) imagination. Still, there is no need to introduce a metaphysically-loaded conception of re-creation (either mode-based or content-based) to ground the distinction between objective and subjective imagination. It is enough that phenomenological contrasts can be drawn between particular cases of objective imagination and particular cases of subjective imagination in various domains. This is exactly how Vender (1984) introduces the distinction in the domain of imagining actions. At this point, our method departs from phenomenology and becomes abductive and specu-

lative. In our view, the best explanation of the relevant phenomenological contrasts is that the imaginings correspond to different kinds of experience in the non-imaginative realm, namely external and internal experiences. We need not rely exclusively on pre-theoretical intuitions. Our taxonomy is indeed grounded in particular phenomenological contrasts, but it is also informed by (controversial) theoretical notions, such as the notion of an external (as opposed to an internal) experience.

2 On the distinction between objective and subjective imagination

Brüggen finds our distinction between objective and subjective imagination “very helpful” ([this collection](#), p. 9), but she is worried about the way we flesh out the distinction. We have already answered one of her worries, which is that our account of the distinction carries a commitment to the mode-based conception of re-creation. As we have suggested, our account is compatible with the alternative, content-based conception. Another worry of Brüggen’s is that it is unclear how our notions of objective and subjective imagination differ from Vendler’s. Brüggen grounds this worry in the fact that our account leaves room for the claim that both objective and subjective imagination always involve the self implicitly ([this collection](#), p. 5).

As far as objective imagination is concerned, our examples certainly suggest that when one objectively (e.g., visually) imagines oneself in an explicit way (e.g., as a rider or as showing a pinched face), one’s imagining can also be implicitly self-involving. This does not mean that the imaginer’s self is involved twice. Here the imaginer’s self is involved only in an explicit way (as we point out all too briefly in the beginning of section 4.1 of our target paper, our definition of implicit self-involvement excludes that the same self that is involved both implicitly and explicitly in a single imagining). The claim that objective imagination is *always* implicitly self-involving does not immediately follow from these examples, but it is admittedly consistent with our account.

Things are more complicated with respect to subjective imagination. We argue that the latter can be either implicitly or explicitly self-involving, although we also acknowledge that the latter is controversial, since it assumes that we can have an internal experience that explicitly represents the self as such. Taking for granted that some subjective imaginings can explicitly involve the self, it is hard to see how they can also be implicitly self-involving. This is so because of the very nature of the re-created internal experience. An internal experience can only be about a (physical or mental) state whose bearer is identical with the bearer of the experience itself. It is not possible to have a proprioceptive experience of another’s body, or to introspect someone else’s mental states. When a subjective imagining re-creates an internal experience that explicitly represents the self (the imaginer’s or someone else’s), the latter cannot but be the self of the re-created experience. Thus the imagining is not implicitly self-involving, according to our definition.

Moreover, even granting Brüggen’s claim that objective and subjective imagination always involve the self implicitly, we do not see how this leads us back to Vendler’s account of the distinction. For us, the key to the distinction is not the distinction between explicit and implicit self-involvement, but rather the distinction between external and internal experiences. Indeed, the latter distinction has to do with aspects of the self, since we have defined an internal experience as being normally *de se*; but, as we have seen, the *de se* nature of internal experiences can be explained independently of whether the self is explicitly or implicitly involved in the relevant imaginings.

Brüggen introduces the notion of an empty point of view as an additional tool for the theory of imagination. For instance, when a subject visually imagines the Panthéon, her imagining involves a perspective that is not occupied by herself or anyone else. In other words, it is not required that there be an observer *in the imaginary world* (the subject can visualize an unseen Panthéon). If this is the right interpretation of Brüggen’s notion of an empty point of view, we already have it in our toolbox. For we claim that

the first-person perspective from which the subject is imagining the Panthéon can remain virtual or counterfactual, in the sense that she is imagining a situation from a spatial perspective that a normally-sighted subject *would* have if she were suitably oriented in the imaginary world.

Brüggen suggests that we could use the notion of an empty point of view to “further sharpen” the distinction between objective and subjective imagination ([this collection](#), p. 6). The idea seems to be that objective imagination always involves an empty point of view, while subjective imagination never does. Let us grant that this idea is broadly correct. We still think that our account of objective and subjective imagination as re-creating external and internal experiences can provide a more fundamental explanation. One might claim that subjective imagination creates more ontological constraints on the imaginary world than objective imagination. A subjective imagining represents a state whose bearer can only be that of the re-created internal experience itself. If such a state is ontologically dependent on a bearer, one cannot imagine the former in a world in which the latter does not exist. Thus, subjective imagination imposes the existence of a self in the imaginary world, whether or not the self in question is explicitly represented. In contrast, since objective imagination re-creates an external experience, one might argue that it is free from the specific constraints of subjective imagination, and need not impose the existence of any self in the imaginary world.

Toward the end of her commentary, Brüggen also suggests that the notion of an empty point of view can help us to distinguish between imaginings and non-imaginative experiences. If we understand her correctly, her suggestion is that in contrast to imaginings, non-imaginative experiences *must* involve an occupied point of view. This is an interesting suggestion, and we do not see why we cannot take it on board. Brüggen thinks otherwise and writes: “Dokic and Arcangeli seem to think that imaginings mirror non-imaginative states with respect to the nature of the point of view involved (again probably partly due to the notion of re-creation)” ([this collection](#), p. 9). However, as de-

tailed above, our account is more neutral and does not carry such a commitment. We do not posit a specific relationship between imaginings and non-imaginative states, but for the sake of argument let us put in a good word for a less neutral view. Even if one claims that imaginings mirror (or simulate) non-imaginative states in the sense that they are dependent on the latter, thus holding an asymmetrical relationship between those kinds of mental states, one is not committed to the conclusion that imaginings mirror every aspect of non-imaginative states (e.g., the nature of the point of view). Further specifications are needed about what precisely is preserved and according to which mapping function ([Arcangeli 2011b](#)).

3 On cognitive imagination

Brüggen is hesitant about our classification of cognitive imaginings as experiential imaginings. Her main reason for being hesitant is not that the notion of cognitive phenomenology is ill-conceived. On the contrary, she is attracted by the view that beliefs have a special phenomenal character. She thinks that cognitive imaginings do not involve an experiential perspective because she construes the notion of experiential perspective quite narrowly, as a spatial egocentric perspective. In our view, Brüggen’s construal of the notion of experiential perspective is too narrow. On this construal, many non-cognitive imaginings turn out to be non-experiential as well. Some cases of sensory imaginings, involving auditory, olfactory, or gustatory imagination, do not always clearly involve a spatial egocentric perspective. Many imaginings that re-create internal experiences (excluding perhaps proprioception) do not involve such a perspective either. For our part, we do not see why the notion of experiential perspective should be restricted to the spatial egocentric case.

4 Conclusion

We have not tried to be exhaustive and answer every point raised in Brüggen’s rich commentary here. But we still hope that we have dealt with her main concerns. Despite the fact that

our minimal notion of re-creation does not introduce a substantial metaphysical relation between the imaginative and the non-imaginative realms, it should be conceived as a placeholder for such a relation. Our taxonomy can then be taken as a starting-point for, and perhaps a constraint on, a full-blooded theory of the ontology of imagination.

References

- Arcangeli, M. (2011a). L'immaginazione ricreativa. *Sistemi intelligenti*, 23 (1), 59-74. [10.1422/34612](https://doi.org/10.1422/34612)
- (2011b). *The imaginative realm and supposition*. Paris, FR: University Paris 6-UPMC PhD Dissertation.
- Brüggen, A. -S. (2015). Imagination and experience: A commentary on Jérôme Dokic and Margherita Arcangeli. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt, a. M.: MIND Group.
- Dokic, J. (2008). Epistemic perspectives on imagination. *Revue Internationale de Philosophie*, 243 (1), 99-118.

On the Eve of Artificial Minds

Chris Eliasmith

I review recent technological, empirical, and theoretical developments related to building sophisticated cognitive machines. I suggest that rapid growth in robotics, brain-like computing, new theories of large-scale functional modeling, and financial resources directed at this goal means that there will soon be a significant increase in the abilities of artificial minds. I propose a specific timeline for this development over the next fifty years and argue for its plausibility. I highlight some barriers to the development of this kind of technology, and discuss the ethical and philosophical consequences of such a development. I conclude that researchers in this field, governments, and corporations must take care to be aware of, and willing to discuss, both the costs and benefits of pursuing the construction of artificial minds.

Keywords

Artificial cognition | Artificial intelligence | Brain modelling | Machine learning | Neuromorphic computing | Robotics | Singularity

Prediction is difficult, especially about the future
– Danish Proverb

1 Introduction

The prediction game is a dangerous one, but that, of course, is what makes it fun. The pitfalls are many: some technologies change exponentially but some don't; completely new inventions, or fundamental limits, might appear at any time; and it can be difficult to say something informative without simply stating the obvious. In short, it's easy to be wrong if you're specific. (Although, it is easy to be right if you're Nostradamus.) Regardless, the purpose of this essay is to play this game. As a consequence, I won't be pursuing technical discussion on the finer points of what a mind is, or how to build one, but rather attempting to paint an abstract portrait of the state of research in fields related to machine intelligence broadly construed. I think the risks of undertak-

ing this kind of prognostication are justified because of the enormous potential impact of a new kind of technology that lies just around the corner. It is a technology we have been dreaming about—and dreading—for hundreds of years. I believe we are on the eve of artificial minds.

In 1958 [Herbert Simon](#) & [Allen Newell](#) claimed that “there are now in the world machines that think” and predicted that it would take ten years for a computer to become world chess champion and write beautiful music ([1958](#), p. 8). Becoming world chess champion took longer, and we still don't have a digital Debussy. More importantly, even when a computer became world chess champion it was not generally seen as the success that Simon and

Author

[Chris Eliasmith](#)
celiasmith@uwaterloo.ca
University of Waterloo
Waterloo, ON, Canada

Commentator

[Daniela Hill](#)
daniela.hill@gmx.net
Johannes Gutenberg-Universität
Mainz, Germany

Editors

[Thomas Metzinger](#)
metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

Newell had expected. This is because the way in which Deep Blue beat Gary Kasparov did not strike many as advancing our understanding of cognition. Instead, it showed that brute force computation, and a lot of careful tweaking by expert chess players, could surpass human performance in a specific, highly circumscribed environment.

Excitement about AI grew again in the 1980s, but was followed by funding cuts and general skepticism in the “AI winter” of the 1990s (Newquist 1994). Maybe we are just stuck in a thirty-year cycle of excitement followed by disappointment, and I am simply expressing the beginning of the next temporary uptick. However, I don’t think this is the case. Instead, I believe that there are qualitative changes in methods, computational platforms, and financial resources that place us in a historically unique position to develop artificial minds. I will discuss each of these in more detail in subsequent sections, but here is a brief overview.

Statistical and brain-like modeling methods are far more mature than they have ever been before. Systems with millions (Garis et al. 2010; Eliasmith et al. 2012) and even tens of millions (Fox 2009) of simulated neurons are suddenly becoming common, and the scale of models is increasing at a rapid rate. In addition, the challenges of controlling a sophisticated, nonlinear body are being met by recent advances in robotics (Cheah et al. 2006; Schaal et al. 2007). These kinds of methodological advances represent a significant shift away from classical approaches to AI (which were largely responsible for the previously unfulfilled promises of AI) to more neurally inspired, and brain-like ones. I believe this change in focus will allow us to succeed where we haven’t before. In short, the conceptual tools and technical methods being developed for studying what I call “biological cognition” (Eliasmith 2013), will make a fundamental difference to our likelihood of success.

Second, there have been closely allied and important advances in the kinds of computational platforms that can be exploited to run these models. So-called “neuromorphic” computing—hardware platforms that perform brain-

style computation—has been rapidly scaling up, with several current projects expected to hit millions (Choudhary et al. 2012) and billions (Khan et al. 2008) of neurons running in real time within the next three to four years. These hardware advances are critical for performing efficient computation capable of realizing brain-like functions embedded in and controlling physical, robotic bodies.

Finally, unprecedented financial resources have been allocated by both public and private groups focusing on basic science and industrial applications. For instance, in February 2013 the European Union announced one billion euros in funding for the Human Brain Project, which focuses on developing a large scale brain model as well as neuromorphic and robotic platforms. A month later, the Obama BRAIN initiative was announced in the United States. This initiative devotes the same level of funding to experimental, technological, and theoretical advances in neuroscience. More recently, there has been a huge amount of private investment:

Google purchased eight robotics and AI companies between Dec 2013 and Jan 2014, including industry leader Boston Dynamics Stunt (2014).

Qualcomm has introduced the Zeroth processor, which is modeled after how a human brain works (Kumar 2013). They demonstrated an Field-Programmable Gate Array (FPGA) mock-up of the chip performing a reinforcement learning task on a robot.

Amazon has recently expressed a desire to provide the Amazon Prime Air service, which will use robotic quadcopters to deliver goods within thirty minutes of their having been ordered (Amazon 2013).

IBM has launched a product based on Watson, which famously beat the best human Jeopardy players (<http://ibm.com/innovation/us/watson/>). The product will provide confidence based responses to natural language queries. It has been opened up to allow developers to use it in a wide variety of applications. They are also developing a neuromorphic platform (Esser et al. 2013).

In addition, there are a growing number of startups that work on brain-inspired computing including Numenta, the Brain Corporation, Vi-

carious, DeepMind (recently purchased by Google for \$400 million) and Applied Brain Research, among many others. In short, I believe there are more dollars being directed at the problem than ever before.

It is primarily these three forces that I believe will allow us to build *convincing* examples of artificial minds in the next fifty years. And, I believe we can do this without necessarily defining what it is that makes a “mind”—even an artificial one. As with many subtle concepts—such as “game,” to use Wittgenstein’s example, or “pornography,” to use Supreme Court Justice Potter Stewart’s example—I suspect we will avoid definitions and rely instead on our sophisticated, but poorly understood, methods of classifying the world around us. In the case of “minds,” these methods will be partly behavioural, partly theoretical, and partly based on judgments of similarity to the familiar. In any case, I do not propose to provide a definition here, but rather to point to reasons why the artifacts we continue to build will become more and more like the natural minds around us. In doing so, I survey recent technological, theoretical, and empirical developments that are important for supporting our progress on this front. I then suggest a timeline over which I expect these developments to take place. Finally, I conclude with what I expect to be the major philosophical and societal impacts on our being able to build artificial minds. As a reminder, I am adopting a somewhat high-level perspective on the behavioural sciences and related technologies in order to make clear where my broad (and likely wrong) predictions are coming from. In addition, if I’m not entirely wrong, I suspect that the practical implications of such developments will prove salient to a broad audience, and so, as researchers in the area, we should consider the consequences of our research.

2 Technological developments

Because I take it that brain-based approaches provide the “difference that makes a difference” between current approaches and traditional AI, here I focus on developments in neuromorphic and robotic technology. Notably, all of the de-

velopments in neuromorphic hardware that I discuss below are inspired by some basic features of neural computation. For instance, all of the neuromorphic approaches use spiking neural networks SNNs to encode and process information. In addition, there is unanimous agreement that biological computation is in orders of magnitude more power efficient than digital computation (Hasler & Marr 2013). Consequently, a central motivation behind exploring these hardware technologies is that they might allow for sophisticated information processing using small amounts of power. This is critical for applications that require the processing to be near the data, such as in robotics and remote sensing. In what follows I begin by providing a sample of several major projects in neuromorphic computing that span the space of current work in the area. I then briefly discuss the current state of high-performance computing and robotics, to identify the roles of the most relevant technologies for developing artificial minds.

To complement its cognitively focused Watson project, IBM has been developing a neuromorphic architecture, a digital model of individual neurons, and a method for programming this architecture (Esser et al. 2013). The architecture itself is called TrueNorth. They argue that the “low-precision, synthetic, simultaneous, pattern-based metaphor of TrueNorth is a fitting complement to the high-precision, analytical, sequential, logic-based metaphor of today’s of von Neumann computers” (Esser et al. 2013, p. 1). TrueNorth has neurons organized into 256 neuron blocks, in which each neuron can receive input from 256 axons. To assist with programming this hardware, IBM has introduced the notion of a “corelet,” which is an abstraction that encapsulates local connectivity in small networks. These act like small programs that can be composed in order to build up more complex functions. To date the demonstrations of the approach have focused on simple, largely feed-forward, standard applications, though across a wide range of methods, including Restricted Boltzmann Machines (RBMs), liquid state machines, Hidden Markov Model (HMMs), and so on. It should be noted that the proposed chip does not yet exist, and

current demonstrations are on detailed simulations of the architecture. However, because it is a digital chip the simulations are highly accurate.

A direct competitor to IBM's approach is the Zeroth neuromorphic chip from Qualcomm. Like IBM, Qualcomm believes that constructing brain-inspired hardware will provide a new paradigm for exploiting the efficiencies of neural computation, targeted at the kind of information processing at which brains excel, but which is extremely challenging for von Neumann approaches. The main difference between these two approaches is that Qualcomm has committed to allowing online learning to take place on the hardware. Consequently, they announced their processor by demonstrating its application in a reinforcement learning paradigm on a real-world robot. They have released videos of the robot maneuvering in an environment and learning to only visit one kind of stimulus (white boxes: <http://www.youtube.com/watch?v=8c1Noq2K96c>). It should again be noted that this is an FPGA simulation of a digital chip that does not yet exist. However, the simulation, like IBM's, is highly accurate.

In the academic sphere, the Spinnaker project at Manchester University has not focused on designing new kinds of chips, but has instead focused on using low-power ARM processors on a massive scale to allow large-scale brain simulations (Khan et al. 2008). As a result, the focus has been on designing approaches to routing that allow for the high bandwidth communication, which underwrites much of the brain's information processing. Simulations on the Spinnaker hardware typically employ spiking neurons, like IBM and Qualcomm, and occasionally allow for learning (Davies et al. 2013), as with Qualcomm's approach. However, even with low power conventional chips, the energy usage is projected to be higher on the Spinnaker platform per neuron. Nevertheless, Spinnaker boards have been used in a wider variety of larger-scale embodied and non-embodied applications. These include simulating place cells, path integration, simple sensory-guided movements, and item classification.

There are also a number of neuromorphic projects that use analog instead of digital implementations of neurons. Analog approaches tend to be several orders of magnitude more power efficient (Hasler & Marr 2013), though also more noisy, unreliable, and subject to process variation (i.e., variations in the hardware due to variability in the size of components on the manufactured chip). These projects include work on the Neurogrid chip at Stanford University (Choudhary et al. 2012), and on a chip at ETH Zürich (Corradi, Eliasmith & Indiveri 2014). The Neurogrid chip has demonstrated larger numbers of simulated neurons—up to a million—while the ETH Zürich chip allows for online learning. More recently, the Neurogrid chip has been used to control a nonlinear, six degree of freedom robotic arm, exhibiting perhaps the most sophisticated information processing from an analog chip to date.

In addition to the above neuromorphic projects, which are focused on cortical simulation, there have been several specialized neuromorphic chips that mimic the information processing of different perceptual systems. For example, the dynamic vision sensor (DVS) artificial retina developed at ETH Zürich performs real-time vision processing that results in a stream of neuron-like spikes (Lichtsteiner et al. 2008). Similarly, an artificial cochlea called AEREAR2 has been developed that generates spikes in response to auditory signals (Li et al. 2012). The development of these and other neuronal sensors makes it possible to build fully embodied spiking neuromorphic systems (Galluppi et al. 2014).

There have also been developments in traditional computing platforms that are important for supporting the construction of models that run on neuromorphic hardware. Testing and debugging large-scale neural models is often much easier with traditional computational platforms such as Graphics Processing Unit (GPUs) and supercomputers. In addition, the development of neuromorphic hardware often relies on simulation of the designs before manufacture. For example, IBM has been testing their TrueNorth architecture with very large-scale simulations that have run up to 500 billion

neurons. These kinds of simulations allow for designs to be stress-tested and fine-tuned before costly production is undertaken. In short, the development of traditional hardware is also an important technological advance that supports the rapid development of more biologically-based approaches to constructing artificial cognitive systems.

A third area of rapid technological development that is critical for successfully realizing artificial minds is the field of robotics. The success of recent methods in robotics have entered public awareness with the creation of the Google car. This self-driving vehicle has successfully navigated hundreds of thousands of miles of urban and rural roadways. Many of the technologies in the car were developed out of DARPA's Grand Challenge to build an autonomous vehicle that would be tested in both urban and rural settings. Due to the success of the first three iterations of the Grand Challenge, DARPA is now funding a challenge to build robots that can be deployed in emergency situations, such as a nuclear meltdown or other disaster.

One of the most impressive humanoid robots to be built for this challenge is the Atlas, constructed by Boston Dynamics. It has twenty-eight degrees of freedom, covering two arms, two legs, a torso, and a head. The robot has been demonstrated walking bipedally, even in extremely challenging environments in which it must use its hands to help navigate and steady itself (<http://www.youtube.com/watch?v=zkBnFPBV3f0>). Several teams in this most recent Grand Challenge have been awarded a copy of Atlas, and have been proceeding to competitively design algorithms to improve its performance.

In fact, there have been a wide variety of significant advances in robotic control algorithms, enabling robots—including quadcopters, wheeled platforms, and humanoid robots—to perform tasks more accurately and more quickly than had previously been possible. This has resulted in one of the first human versus robot dexterity competitions being recently announced. Just as IBM pitted Watson against human Jeopardy champions, Kuka has pitted its high-speed arm against the human

ping-pong champion Timo Boll (http://www.youtube.com/watch?v=_mbdtupCbc4). Despite the somewhat disappointing outcome, this kind of competition would not have been thought possible a mere five years ago (Ackerman 2014).

These three areas of technological development—neuromorphics, high-performance conventional computing, and robotics—are progressing at an incredibly rapid pace. And, more importantly, their convergence will allow a new class of artificial agents to be built. That is, agents that can begin processing information at very similar speeds and support very similar skills to those we observe in the animal kingdom. It is perhaps important to emphasize that my purpose here is predictive. I am not claiming that current technologies are sufficient for building a new kind of artificial mind, but rather that they lay the foundations, and are progressing at a sufficient rate to make it reasonable to expect that the sophistication, adaptability, flexibility, and robustness of artificial minds will rapidly approach those of the human mind. We might again worry that it will be difficult to measure such progress, but I would suggest that progress will be made along many dimensions simultaneously, so picking nearly any of dimensions will result in some measurable improvement. In general, multi-dimensional similarity judgements are likely to result in “I’ll know it when I see it” kinds of reactions to classifying complicated examples. This may be derided by some as “hand-wavy”, but it might also be a simple acknowledgement that “mindedness” is complex. I would like to be clear that my claims about approaching human mindful behaviour are to be taken as applying to the vast majority of the many measures we use for identifying minds.

3 Theoretical developments

Along with these technological developments there have been a series of theoretical developments that are critical for building large-scale artificial agents. Some have argued that theoretical developments are not that important: suggesting that standard back propagation at a sufficiently large scale is enough to capture

complex perceptual processing (Krizhevsky et al. 2012). That is, building brain-like models is more a matter of getting a sufficiently large computer with enough parameters and neurons than it is of discovering some new principles about how brains function. If this is true, then the technological developments that I pointed to in the previous section may be sufficient for scaling to sophisticated cognitive agents. However, I am not convinced that this is the case.

As a result, I think that theoretical developments in deep learning, nonlinear adaptive control, high dimensional brain-like computing, and biological cognition *combined* will be important to support continued advances in understanding how the mind works. For instance, deep networks continue to achieve state-of-the-art results in a wide variety of perception-like processing challenges (http://rodrigob.github.io/are_we_there_yet/build/classification_datasets_results.html#43494641522d3130). And while deep networks have traditionally been used for static processing, such as an image classification or document classification, there has been a recent, concerted move to use them to model more dynamic perceptual tasks as well (Graves et al. 2013). In essence, deep networks are one among many techniques for modeling the statistics of time varying signals, a skill central to animal cognition.

However, animals are also incredibly adept at *controlling* nonlinear dynamical systems, including their bodies. That is, biological brains can *generate* time varying signals that allow successful and sophisticated interactions with their environment through their body. Critically, there have been a variety of important theoretical advances in nonlinear and adaptive control theory as well. New methods for solving difficult optimal control problems have been discovered through careful study of biological motor control (Schaal et al. 2007; Todorov 2008). In addition, advances in hierarchical control allow for real-time computation of difficult inverse kinematics problems on a laptop (Khatib 1987). And, finally, important advances in adaptive control allow for the automatic learning of both kinematic and dynamic models even in highly nonlinear and

high dimensional control spaces (Cheah et al. 2006).

Concurrently with these more abstract characterizations of brain function there have been theoretical developments in neuroscience that have deepened our understanding of how biological neural networks may perform sophisticated information processing. Work using the Neural Engineering Framework (NEF) has resulted in a wide variety of spiking neural models that mirror data recorded from biological systems (Eliasmith & Anderson 1999, 2003). In addition, the closely related liquid computing (Maass et al. 2002) and FORCE learning (Sussillo & Abbott 2009) paradigms have been successfully exploited by a number of researchers to generate interesting dynamical systems that often closely mirror biological data. Together these kinds of methods provide quantitative characterizations of the computational power available in biologically plausible neural networks. Such developments are crucial for exploiting neuromorphic approaches to building brain-like hardware. And they suggest ways of testing some of the more abstract perceptual and control ideas in real-world, brain-like implementations.

Interestingly, several authors have suggested that difficult perceptual and control problems are in fact mathematical duals of one another (Todorov 2009; Eliasmith 2013). This means that there are deep theoretical connections between perception and motor control. This realization points to a need to think hard about how diverse aspects of brain function can be integrated into single, large-scale models. This has been a major focus of research in my lab recently. One result of this focus is Spaun, currently the world's largest functional brain model. This model incorporates deep networks, recent control methods, and the NEF to perform eight different perceptual, motor, and cognitive tasks (Eliasmith et al. 2012). Importantly, this is not a one-off model, but rather a single example among many that employs a general architecture intended to directly address integrated biological cognition (Eliasmith 2013). Currently, the most challenging constraints for running models like Spaun are technological—computers are not fast enough. However, the

neuromorphic technologies mentioned previously should soon remove these constraints. So, in some sense, theory currently outstrips application: we have individually tested several critical assumptions of the model and shown that they scale well (Crawford et al. 2013), but we are not yet able to integrate full-scale versions of the components due to limitations in current computational resources.

Taken together, I believe that these recent theoretical developments demonstrate that we have a roadmap for how to approach the problem of building sophisticated models of biological cognition. No doubt not all of the methods we need are currently available, but it is not evident that there are any major conceptual roadblocks to building a cognitive system that rivals the flexibility, adaptability, and robustness of those found in nature. I believe this is a unique historical position. In the heyday of the symbolic approach to AI there were detractors who said that the perceptual problems solved easily by biological systems would be a challenge for the symbolic approach (Norman 1986; Rumelhart 1989). They were correct. In the heyday of connectionism there were detractors who said that standard approaches to artificial neural networks would not be able to solve difficult planning or syntactic processing problems (Pinker & Prince 1988; Fodor & Pylyshyn 1988; Jackendoff 2002). They were correct. In the heyday of statistical machine learning approaches (a heyday we are still in) there are detractors who say that mountains of data are not sufficient for solving the kinds of problems faced by biological cognitive systems (Marcus 2013). They are probably correct. However, as many of the insights of these various approaches are combined with control theory, integrated into models able to do efficient syntactic and semantic processing with neural networks, and, in general, become conceptually *unified* (Eliasmith 2013), it is less and less obvious what might be missing from our characterization of biological cognition.

4 Empirical developments

One thing that might be missing is, simply, knowledge. We have many questions about how

real biological systems work that remain unanswered. Of course, complete knowledge of natural systems is not a prerequisite for building nearly functionally equivalent systems (see e.g., flight). However, I believe our understanding of natural cognitive systems will continue to play an important role in deciding what kinds of algorithms are worth pursuing as we build more sophisticated artificial agents.

Fortunately, on this front there have been two announcements of significant resources dedicated to improving our knowledge of the brain, which I mentioned in the introduction. One is from the EU and the other from the US. Each are investing over \$1 billion in generating the kind of data needed to fill gaps in our understanding of how brains function. The EU's Human Brain Project (HBP) includes two central subprojects aimed at gathering mouse and human brain data to complement the large-scale models being built within the project. These subprojects will focus on genetic, cellular, vascular, and overall organizational data to complement the large-scale projects of this type already available (such as the Allen Brain Atlas, <http://www.brain-map.org/>). One central goal of these subprojects is to clarify the relationship between the mouse (which is highly experimentally accessible) and human subjects.

The American "brain research through advancing innovative neurotechnologies" (BRAIN) initiative is even more directly focused on large-scale gathering of neural data. Its purpose is to accelerate technologies to provide large-scale dynamic information about the brain that demonstrates how both single runs and larger neural circuits operate. Its explicit goal is to "fill major gaps in our current knowledge" (<http://www.nih.gov/science/brain/>). It is a natural complement to the human connectome project, which has been mapping the structure of the human brain on a large-scale (<http://www.humanconnectomeproject.org/>). Even though it is not yet clear exactly what information will be provided by the BRAIN initiative, it is clear that significant resources are being put into developing technologies that draw on nanoscience, informatics, engineering, and other fields to measure the brain at a level of detail and scale not previously possible.

Table 1

Within (years)	Von Neumann (real-time neurons)	Neuromorphic (real-time neurons)	Behaviours
5	10^8	10^7	constrained environment navigation; uncluttered vision/audition recognition; simple language understanding; slow, robust motor control
10	10^9	10^8	good open world large-scale navigation; learn simple few-step cognitive tasks; rapid single limb motor control; general, shallow semantics and syntax
15	10^{10}	10^9	literal natural language (4-5 year old equivalent); learn new many step tasks; near human quality perceptual skills (categorization and recognition in arbitrary, moving scenes)
25	10^{11}	10^{11}	arbitrary autonomous navigation; highly tunable cognitive system for task specialization, generally exceeding human ability on gross manual tasks, basic problem solving, etc.; human quality perception
50	10^{13}	10^{14}	human level full body agility; above average IQ; nuanced natural language; better than human perception

While both of these projects are just over a year old, they have both garnered international attention and been rewarded with sufficient funding to ensure a good measure of success. Consequently, it is likely that as we build more sophisticated models of brain function, and as we discover where our greatest areas of ignorance lay, we will be able to turn to the methods developed by these projects to rapidly gain critical information and continue improving our models. In short, I believe that there is a confluence of technological, theoretical, and empirical developments that will allow for bootstrapping detailed functional models of the brain. It is precisely these kinds of models that I expect will lead to the most convincing embodiments of artificial cognition that we have ever seen—I am even willing to suggest that their sophistication will rival those of natural cognitive systems.

5 A future timeline

Until this point I have been mustering evidence that there will soon be significant improvements in our ability to construct artificial cognitive agents. However, I have not been very specific about timing. The purpose of this section is to provide more quantification on the speed of development in the field.

In Table 1, the first column specifies the timeframe, the second suggests the number of neurons that will be simulatable in real-time on standard hardware, the third suggests the number of neurons that will be simulatable in real-time on neuromorphic hardware, and the last identifies relevant achievable behaviours within that timeframe.

I believe that several of the computational technologies I have mentioned, as well as empir-

ical methods for gathering evidence, are on an exponential trajectory by relevant measures (e.g., number of neurons per chip, number of neurons recorded [Stevenson & Kording 2011](#)). On the technological side, if we assume a doubling every eighteen months, this is equivalent to an increase of about one order of magnitude every five years. I should also note that I am assuming that real-time simulation of neurons will be embedded in an interactive, real-world environment, and that the neuron count is for the whole system (not a single chip). For context, it is worth remembering that the human brain has about 10^{11} neurons, though they are more computationally sophisticated than those typically simulated in hardware.

Another caveat is that it is likely that large-scale simulations on a digital Von Neumann architecture will hit a power barrier, which makes it likely that the suggested scaling could be achieved, but will be cost-prohibitive in fifty years. Consequently, a neuromorphic alternative is most likely to be the standard implementational substrate of artificial agents.

Finally, the behavioural characterizations I am giving are with a view to functions necessary for creating a convincing artificial mind in an artificial body. Consequently, my comments generally address perceptual, motor, and cognitive skills relevant to reproducing human-like abilities.

6 Consequences for philosophy

So suppose that, fifty years hence, we have developed an understanding of cognitive systems that allows us to build artificial systems that are on par with, or, if we see fit, surpass the abilities of an average person. Suppose, that is, that we can build artificial agents that can move, react, adapt, and think much like human beings. What consequences, if any, would this have for our theoretical questions about cognition? I take these questions to largely be in the domain of philosophy of mind. In this section I consider several central issues in philosophy of mind and discuss what sorts of consequences I take building a human-like artificial agent to have for them.

Being a philosopher, I am certain that, for any contemporary problem we consider, at least some subset of those who have a committed opinion about that problem will not admit that any amount of technical advance can “solve” it. I suspect, however, that their opinions may end up carrying about as much weight as a modern-day vitalist. To take one easy example, let us think for a moment about contemporary dualism. Some contemporary dualists hold that even if we had a complete understanding of how the brain functions, we would be no closer to solving the “hard problem” of consciousness ([Chalmers 1996](#)). The “hard problem” is the problem of explaining how subjective experience comes from neural activity. That is, how the phenomenal experiences we know from a first-person perspective can be accounted for by third-person physicalist approaches to understanding the mind. If indeed we have constructed artificial agents that behave much like people, share a wide variety of internal states with people, are fully empirically accessible, and report experiences like people, it is not obvious to what extent this problem will not have been solved. Philosophers who are committed to the notion that no amount of empirical knowledge will solve the problem will of course dismiss such an accomplishment on the strength of their intuitions. I suspect, however, that when most people are actually confronted with such an agent—one they can interrogate to their heart’s content and one about which they can have complete knowledge of its functioning—it will seem odd indeed to suppose that we cannot explain how its subjective experience is generated. I suspect it will seem as odd as someone nowadays claiming that we cannot expect to explain how life is generated despite our current understanding of biochemistry. Another way to put this is that the “strong intuitions” of contemporary dualists will hold little plausibility in the face of actually existing, convincing artificial agents, and so, I suspect, they will become even more of a rarity.

I refer to this example as “easy” because the central reasons for rejecting dualism are only strengthened, not *generated*, by the existence of sophisticated artificial minds. That is,

good arguments against the dualist view are more or less independent of the current state of constructing agents (although the existence of such agents will likely sway intuitions). However, other philosophical conundrums, like Searle's famous Chinese room (1980), have responses that depend fairly explicitly on our ability to construct artificial agents. In particular, the "systems reply" suggests that a sufficiently complex system will have the same intentional states as a biological cognitive system. For those who think that this is a good rejection of Searle's strong intentionalist views, having systems that meet all the requirements of their currently hypothetical agents would provide strong empirical evidence consistent with their position. Of course, the existence of such artificial agents is unlikely to convince those, like Searle, who believe that there is some fundamental property of biology that allows intentionality to gain a foothold. But it does make such a position seem that much more tenuous if every means of measuring intentionality produces similar measurements across non-biological and biological agents. In any case, the realization of the systems reply does ultimately depend on our ability to construct sufficiently sophisticated artificial agents. And I am suggesting that such agents are likely to be available in the next fifty years.

More immediately, I suspect we will be able to make significant headway on several problems that have been traditionally considered philosophical *before* we reach the fifty-year mark. For example, the frame problem—i.e., the problem of knowing what representational states to update in a dynamic environment—is one that contemporary methods, like control theory and machine learning, struggle with much less than classical methods. Because the dynamics of the environment are explicitly included in the world-model being exploited by such control theoretic and statistical methods, updating state representations naturally includes the kinds of expectations that caused such problems for symbolic approaches.

Similarly, explicit quantitative solutions are suggested for the symbol-grounding problem through integrated models that incorporate as-

pects of both statistical perceptual processing and syntactic manipulation. Even in simple models, like Spaun, it is clear how the symbols for digits that are syntactically manipulated are related to inputs coming from the external world (Eliasmith 2013). And it is clear how those same symbols can play a role in driving the model's body to express its knowledge about those representations. As a result, the tasks that Spaun can undertake demonstrate both conceptual knowledge, through the symbol-like relationships between numbers (e.g., in the counting task), and perceptual knowledge, through categorization and the ability to drive its motor system to reproduce visual properties (e.g., in the copy-drawing task).

In some cases, rather than resolving philosophical debates, the advent of sophisticated artificial agents is likely to make these debates far more empirically grounded. These include debates about the nature of concepts, conceptual change, and functionalism, among others. However these debates turn out, it seems clear that having an engineered, working system that can generate behaviour as sophisticated as that that gave rise to these theoretical ideas in the first place will allow a systematic investigation of their appropriate application. After all, there are few, if any, limits on the empirical information we can garner from such constructed systems. In addition, our having built the system explicitly makes it unlikely that we would be unaware of some "critical element" essential in generating the observed behaviours.

Even without such a working system, I believe that there are already hints as to how these debates are likely to be resolved, given the theoretical approaches I highlighted earlier. For instance, I suspect that we will find that concepts are explained by a combination of vector space representations and a restricted class of dynamic processes defined over those spaces (Eliasmith 2013). Similarly, quantifying the adaptive nature of those representations and processes will indicate the nature of mechanisms of conceptual change in individuals (Thagard 2014). In addition, functionalism will probably seem too crude a hypothesis given a detailed understanding of how to build a wide variety of

artificial minds. Perhaps a kind of “functionalism with error bars” will take its place, providing a useful means of talking about degrees of functional similarity and allowing a quantification of functional characterizations of complex systems. Consequently, suggestions about which functions are or are not necessary for “mindedness” can be empirically tested through explicit implementation and experimentation. This will not solve the problem of mapping experimental results to conceptual claims (a problem we currently face when considering non-human and even some human subjects), but it will make functionalism as empirically accessible as seems plausible.

In addition to these philosophical issues that may undergo reconceptualization with the construction of artificial minds, there are others that are bound to become more vexing. For example, the breadth of application of ethical theory may, for the first time, reach to engineered devices. If, after all, we have built artificial minds capable of understanding their place in the universe, it seems likely we will have to worry about the possibility of their suffering (Metzinger 2013). It does not seem that understanding how such devices work, or having explicitly built them, will be sufficient for dismissing them as having no moral status. While current theories of non-human ethics have been developed, it is not clear how much or little theories of non-biological ethics will be able to borrow from them.

I suspect that the complexities introduced to ethical theory will go beyond adding a new category of potential application. Because artificial minds will be designed, they may be designed to make what have traditionally been morally objectionable inter-mind relationships seem less problematic. Consider, for instance, a robot that is designed to gain maximal self-fulfillment out of providing service to people. That is, unlike any biological species of which we are aware, these robots place service to humans above all else. Is a slave-like relationship between humans and these minds still wrong in such an instance? Whatever our analysis of why slavery is wrong, it seems likely that we will be able to design artificial minds that bypass that

analysis. This is a unique quandary because while it is currently possible for certain individuals to claim to have such slave-aligned goals, it is always possible to argue that they are simply mistaken in their personal psychological analysis. In the case of minds whose psychology is designed in a known manner, however, the having of such goals will at least seem much more genuine. This is only one among many new kinds of challenges that ethical theory will face with the development of sophisticated artificial agents (Metzinger 2013).

I do not take this surely unreasonably brief discussion of any of these subtle philosophical issues to do justice to them. My main purpose here is to provide a few example instances of how the technological developments discussed earlier are likely to affect our theoretical inquiry. On some occasions such developments will lead to strengthening already common intuitions; on others they may provide deep empirical access to closely related issues; and on still other occasions these developments will serve to make complex issues even more so.

7 The good and the bad

As with the development of many technologies—cars, electricity, nuclear power—the construction of artificial minds is likely to have both negative and positive impacts. However, there is a sense in which building *minds* is much more fraught than these other technologies. We may, after all, build agents that are themselves capable of immorality. Presumably we would much prefer to build Commander Data than to build HAL or the Terminator. But how to do this is by no means obvious. There have been several interesting suggestions as to how this might be accomplished, perhaps most notably from Isaac Asimov in his entertaining and thought-provoking exploration of the three laws of robotics. For my purposes, however, I will sidestep this issue—not because it is not important, but because more immediate concerns arise from considering the development of these agents from a technological perspective. Let me then focus on the more immediately pressing consequences of constructing intelligent machines.

The rapid development of technologies related to artificial intelligence has not escaped the notice of governments around the world. One of the primary concerns for governments is the potentially massive changes in the nature of the economy that may result from an increase in automatization. It has recently been suggested that almost half of the jobs in the United States are likely to be computerized in the next twenty years (Rutkin 2013). The US Bureau of Labor and Statistics regularly publishes articles on the significant consequence of automation for the labour force in their journal *Monthly Labor Review* (Goodman 1996; Plewes 1990). This work suggests that greater automatization of jobs may cause standard measures of productivity and output to increase, while still increasing unemployment.

Similar interest in the economic and social impacts of automatization is evident in many other countries. For instance, Policy Horizons Canada is a think-tank that works for the Canadian government, which has published work on the effects of increasing automatization and the future of the economy (Arshad 2012). Soon after the publication of our recent work on Spaun, I was contacted by this group to discuss the impact of Spaun and related technologies. It was clear from our discussion that machine learning, automated control, robotics, and so on are of great interest to those who have to plan for the future, namely our governments and policy makers (Padbury et al. 2014).

This is not surprising. A recent McKinsey report suggests that these highly disruptive technologies are likely to have an economic value of about \$18 trillion by 2025 (Manyika et al. 2013). It is also clear from the majority of analyses, that lower-paid jobs will be the first affected, and that the benefits will accrue to those who can afford what will initially be expensive technologies. Every expectation, then, is that automatization will exacerbate the already large and growing divide between rich and poor (Malone 2014; “The Future of Jobs: The On-rushing Wave” 2014). Being armed with this knowledge now means that individuals, governments, and corporations can support progressive policies to mitigate these kinds of potentially

problematic societal shifts (Padbury et al. 2014).

Indeed, many of the benefits of automatization may help alleviate the potential downsides. Automatization has already had significant impact on the growth of new technology, both speeding up the process of development and making new technology cheaper. The human genome project was a success largely because of the automatization of the sequencing process. Similarly, many aspects of drug discovery can be automatized by using advanced computational techniques (Leung et al. 2013). Automatization of more intelligent behaviour than simply generating and sifting through data is likely to have an even greater impact on the advancement of science and engineering. This may lead more quickly to cleaner and cheaper energy, advances in manufacturing, decreases in the cost and access to advanced technologies, and other societal benefits.

As a consequence, manufacturing is likely to become safer—a trend already seen in areas of manufacturing that employ large numbers of robots (Robertson et al. 2005). At the same time, additional safety considerations come into play as robotic and human workspaces themselves begin to interact. This concern has resulted in a significant focus in robotics on compliant robots. Compliant robots are those that have “soft” environmental interactions, often implemented by including real or virtual springs on the robotic platform. As a result, control becomes more difficult, but interactions become much safer, since the robotic system does not rigidly go to a target position even if there is an unexpected obstacle (e.g., a person) in the way.

As the workplace continues to become one where human and automated systems co-operate, additional concerns may arise as to what kinds of human-machine relationships employers should be permitted to demand. Will employees have the right not to work with certain kinds of technology? Will employers still have to provide jobs to employees who refuse certain work situations? These questions touch on many of the same subjects highlighted in the previous section regarding

the ethical challenges that will be raised as we develop more and more sophisticated artificial minds.

Finally, much has been made of the possibility that the automatization of technological advancement could eventually result in machines designing themselves more effectively than humans can. This idea has captured the public imagination, and the point in time where this occurs is now broadly known as “The Singularity,” a term first introduced by von Neumann (Ulam 1958). Given the vast variety of functions that machines are built to perform, it seems highly unlikely that there will be anything analogous to a mathematical singularity—a clearly defined, discontinuous point—after which machines will be superior to humans. As with most things, such a shift, if it occurs, is likely to be gradual. Indeed, the earlier timeline is one suggestion for how such a gradual shift might occur. Machines are already used in many aspects of design, for performing optimizations that would not be possible without them. Machines are also already much better at many functions than people: most obviously mechanical functions, but more recently cognitive ones, like playing chess and answering trivia questions in certain circumstances.

Because the advancement of intelligent machines is likely to continue to be a smooth, continuous one (even if exponential at times), we will likely remain in a position to make informed decisions about what they are permitted to do. As with members of a strictly human society, we do not tolerate arbitrary behaviour simply because such behaviour is possible. If anything, we will be in a *better* position to specify appropriate behaviour in machines than we are in the case of our human peers. Perhaps we will need laws and other societal controls for determining forbidden or tolerable behaviour. Perhaps some people and machines will choose to ignore those laws. But, as a society, it is likely that we will enforce these behavioural constraints the same way we do now—with publicly sanctioned agencies that act on behalf of society. In short, the dystopian predictions we often see that revolve around the development of intelligent robots seem no more or less likely

because of the robots. Challenges to societal stability are nothing new: war, hunger, poverty, weather are constant destabilizing forces. Artificial minds are likely to introduce another force, but one that may be just as likely to be stabilizing as problematic.

Unsurprisingly, like many other technological changes, the development of artificial minds will bring with it both costs and benefits. It may even be the case that deciding what is a cost and what is a benefit is not straightforward. If indeed many jobs become automated, it would be unsurprising if the average working week becomes shorter. As a result, a large number of people may have much more recreational time than has been typical in recent history. This may seem like a clear benefit, as many of us look forward to holidays and time off work. However, it has been argued that fulfilling work is a central to human happiness (Thagard 2010). Consequently, overly limited or unchallenging work may end up being a significant cost of automation.

As good evidence for costs and benefits becomes available, decision-makers will be faced with the challenge of determining what the appropriate roles of artificial minds should be. These roles will no doubt evolve as technologies change, but there is little reason to presume that unmanageable upheavals or “inflection points” will be the result of artificial minds being developed. While we, as a society, must be aware of, and prepared for, being faced with new kinds of ethical dilemmas, this has been a regular occurrence during the technological developments of the last several hundred years. Perhaps the greatest challenges will arise because of the significant wealth imbalances that may be exacerbated by limited access to more intelligent machines.

8 Conclusion

I have argued that we are at a unique point in the development of technologies that are critical to the realization of artificial minds. I have even gone so far as to predict that human-level intelligence and physical ability will be achieved in about fifty years. I suspect that for many famil-

iar with the history of artificial intelligence such predictions will be easily dismissed. Did we not have such predictions over fifty years ago? Some have suggested that the singularity will occur by 2030 (Vinge 1993), others by 2045 (Kurzweil 2005). There were suggestions and significant financial speculation that AI would change the world economy in the 1990s, but this never happened. Why would we expect anything to be different this time around?

In short, my answer is encapsulated by the specific technological, theoretical, and empirical developments I have described above. I believe that they address the central limitations of previous approaches to artificial cognition, and are significantly more mature than is generally appreciated. In addition, the limitations they address—such as power consumption, computational scaling, control of nonlinear dynamics, and integrating large-scale neural systems—have been more central to prior failures than many have realized. Furthermore, the financial resources being directed towards the challenge of building artificial minds is unprecedented. High-tech companies, including Google, IBM, and Qualcomm have invested billions of dollars in machine intelligence. In addition, funding agencies including DARPA (Defense Advanced Research Projects Agency), EU-IST (European Union—Information Society Technologies), IARPA (Intelligence Advanced Research Projects Agency), ONR (Office of Naval Research), and AFOSR (Air Force Office of Scientific Research) have contributed a similar or greater amount of financial support across a wide range of projects focused on brain-inspired computing. And the two special billion dollar initiatives from the US and EU will serve to further deepen our understanding of biological cognition, which has, and will continue, to inspire builders of artificial minds.

While I believe that the alignment of these forces will serve to underpin unprecedented advances in our understanding of biological cognition, there are several challenges to achieving the timeline I suggest above. For one, robotic actuators are still far behind the efficiency and speeds found in nature. There will no doubt be advances in materials science that will help overcome these limitations, but how long that will take is not yet

clear. Similarly, sensors on the scale and precision of those available from nature are not yet available. This is less true for vision and audition, but definitely the case for proprioception and touch. The latter are essential for fluid, rapid motion control. It also remains to be seen how well our theoretical methods for integrating complex systems will scale. This will only become clear as we attempt to construct more and more sophisticated systems. This is perhaps the most fragile aspect of my prediction: expecting to solve difficult algorithmic and integration problems. And, of course, there are myriad other possible ways in which I may have underestimated the complexity of biological cognition: maybe glial cells are performing critical computations; maybe we need to describe genetic transcription processes in detail to capture learning; maybe we need to delve to the quantum level to get the explanations we need—but I am doubtful (Litt et al. 2006).

Perhaps it goes without saying that, all things considered, I believe the timeline I propose is a plausible one.¹ This, of course, is predicated on there being the societal and political will to allow the development of artificial minds to proceed. No doubt researchers in this field need to be responsive to public concerns about the specific uses to which such technology might be put. It will be important to remain open, self-critical, and self-regulating as artificial minds become more and more capable. We must usher in these technologies with care, fully cogniscent of, and willing to discuss, both their costs and their benefits.

Acknowledgements

I wish to express special thanks to two anonymous reviewers for their helpful feedback. Many of the ideas given here were developed in discussion with members of the CNRG Lab, participants at the Telluride workshops, and my collaborators on ONR grant N000141310419 (PIs: Kwabena Boahen and Rajit Manohar). This work was also funded by AFOSR grant FA8655-13-1-3084, Canada Research Chairs, and NSERC Discovery grant 261453.

¹ It is quite different from that proposed by the HBP, for example. For further discussion of the differences in perspective between the HBP and my lab's work, see Eliasmith & Trujillo (2013).

References

- Ackerman, E. (2014). Robots playing ping pong: What's real, what's not? *IEEE Spectrum*
- Amazon (2013). Amazon Prime Air.
- Arshad, I. (2012). People and machines: Competitors or collaborators in the emerging world of work?
- Chalmers, D. (1996). *The conscious mind: In search of a fundamental theory*. Oxford, UK: Oxford University Press.
- Cheah, C. C., Liu, C. & Slotine, J. J. E. (2006). Adaptive tracking control for robots with unknown kinematic and dynamic properties. *The International Journal of Robotics Research*, 25 (3), 283-296. SAGE Publications.
- Choudhary, S., Sloan, S., Fok, S., Neckar, A., Trautmann, E., Gao, P., Stewart, T., Eliasmith, C. & Boahen, K. (2012). Silicon neurons that compute. In A. E. P. Villa, W. Duch, P. Érdi, F. Masulli and G. Palm (Eds.) *Artificial neural networks and machine learning - ICANN 2012* (pp. 121-128). Berlin, GER: Springer. [10.1007/978-3-642-33269-2_16](https://doi.org/10.1007/978-3-642-33269-2_16)
- Corradi, F., Eliasmith, C. & Indiveri, G. (2014). Mapping arbitrary mathematical functions and dynamical systems to neuromorphic VLSI circuits for spike-based neural computation. *IEEE International Symposium on Circuits and Systems (ISCAS) 2014* (pp. 269-272). IEEE. [10.1109/ISCAS.2014.6865117](https://doi.org/10.1109/ISCAS.2014.6865117)
- Crawford, E., Gingerich, M. & Eliasmith, C. (2013). Biologically plausible, human-scale knowledge representation. In M. Knauff, M. Pauen, N. Sebanz and I. Wachsmuth (Eds.) *35th Annual Conference of the Cognitive Science Society* (pp. 412-417). Austin, TX: Cognitive Science Society.
- Davies, S., Stewart, T., Eliasmith, C. & Furber, S. (2013). Spike-based learning of transfer functions with the SpiNNaker neuromimetic simulator. *The 2013 International Joint Conference on Neural Networks (IJCNN)* (pp. 1832-1839). IEEE. [10.1109/IJCNN.2013.6706962](https://doi.org/10.1109/IJCNN.2013.6706962)
- Eliasmith, C. (2013). *How to build a brain: A neural architecture for biological cognition*. New York, NY: Oxford University Press.
- Eliasmith, C. & Anderson, C.H. (1999). Developing and applying a toolkit from a general neurocomputational framework. *Neurocomputing*, 26-27 (0), 1013-1018. [10.1016/S0925-2312\(99\)00098-3](https://doi.org/10.1016/S0925-2312(99)00098-3)
- (2003). *Neural engineering: Computation, representation and dynamics in neurobiological systems*. Cambridge, MA: MIT Press.
- Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y. & Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science*, 338 (6111), 1202-1205. [10.1126/science.1225266](https://doi.org/10.1126/science.1225266)
- Eliasmith, C. & Trujillo, O. (2013). The use and abuse of large-scale brain models. *Current Opinion in Neurobiology*, 25, 1-6. [10.1016/j.conb.2013.09.009](https://doi.org/10.1016/j.conb.2013.09.009)
- Esser, S. K., Andreopoulos, A., Appuswamy, R., Datta, P., Barch, D., Amir, A., Arthur, J., Cassidy, A., Flickner, M., Merolla, P., Chandra, S., Basilico, N., Carpin, S., Zimmerman, T., Zee, F., Alvarez-Icaza, R., Kusnitz, J. A., Wong, T. M., Risk, W. P., McQuinn, E., Nayak, T. K., Singh, R. & Modha, D. S. (2013). Cognitive computing systems: Algorithms and applications for networks of neurosynaptic cores. *The 2013 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-10). IEEE. [10.1109/IJCNN.2013.6706746](https://doi.org/10.1109/IJCNN.2013.6706746)
- Fodor, J. A. & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28 (1-2), 3-71. [10.1016/0010-0277\(88\)90031-5](https://doi.org/10.1016/0010-0277(88)90031-5)
- Fox, D. (2009). IBM reveals the biggest artificial brain of all time. *Popular Mechanics*
- Galluppi, F., Denk, C., Meiner, M. C., Stewart, T., Plana, L. A., Eliasmith, C., Furber, S. & Conradt, J. (2014). Event-based neural computing on an autonomous mobile platform. *Proceedings of IEEE international conference on robotics and automation (ICRA)*. IEEE.
- De Garis, H., Shuo, C., Goertzel, B. & Ruiting, L. (2010). A world survey of artificial brain projects, Part I: Large-scale brain simulations. *Neurocomputing*, 74 (1-3), 3-29. [10.1016/j.neucom.2010.08.004](https://doi.org/10.1016/j.neucom.2010.08.004)
- Goodman, W. C. (1996). Software and engineering industries: Threatened by technological change? *Monthly Labor Review*, 119 (8), 37-45. Bureau of Labor Statistics, U.S. Department of Labor. [10.2307/41844604](https://doi.org/10.2307/41844604)
- Graves, A., Mohamed, A. & Hinton, G. E. (2013). speech recognition with deep recurrent neural networks. *IEEE international conference on acoustic speech and signal processing (ICASSP)* (pp. 6645-6649). Vancouver, Canada: IEEE.
- Hasler, J. & Marr, H. B. (2013). Finding a roadmap to achieve large neuromorphic hardware systems. *Frontiers in Neuroscience*, 7 (118), 1-29. [10.3389/fnins.2013.00118](https://doi.org/10.3389/fnins.2013.00118)
- Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford, UK: Oxford University Press.
- Khan, M. M., Lester, D. R., Plana, L. A., Rast, A., Jin, X., Painkras, E. & Furber, S. B. (2008). *SpiNNaker*:

- Mapping neural networks onto a massively-parallel chip multiprocessor*. IEEE. (pp. 2849-2856).
[10.1109/IJCNN.2008.4634199](https://doi.org/10.1109/IJCNN.2008.4634199)
- Khatib, O. (1987). A unified approach for motion and force control of robot manipulators: The operational space formulation. *IEEE Journal of Robotics and Automation*, 3 (1), 43-53.
- Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger (Eds.) *Advances in Neural Information Processing Systems 25* (p. 4).
- Kumar, S. (2013). Introducing qualcomm zeroth processors: Brain-inspired computing. <http://www.qualcomm.com/media/blog/2013/10/10/introducing-qualcomm-zeroth-processors-brain-inspired-computing>
- Kurzweil, R. (2005). *The singularity is near*. New York, NY: Penguin Books.
- Leung, E. L., Cao, Z., Jiang, Z., Zhou, H. & Liu, L. (2013). Network-based drug discovery by integrating systems biology and computational technologies. *Briefings in bioinformatics*, 14 (4), 491-505. Oxford, UK: Oxford University Press.
- Lichtsteiner, P., Posch, C. & Delbruck, T. (2008). Temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43 (2), 566-576.
- Li, C., Delbruck, T. & Liu, S. (2012). Real-time speaker identification using the AEREAR2 event-based silicon cochlea. *2012 IEEE International Symposium on Circuits and Systems* (pp. 1159-1162).
[10.1109/ISCAS.2012.6271438](https://doi.org/10.1109/ISCAS.2012.6271438)
- Litt, A., Eliasmith, C., Kroon, F., Weinstein, S. & Thagard, P. (2006). Is the brain a quantum computer? *Cognitive Science*, 30 (3), 593-603.
- Maass, W., Natschläger, T. & Markram, H. (2002). Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural computation*, 14 (11), 2531-2560. Cambridge, MA: MIT Press.
- Malone, P. (2014). Wealthy need to share the spoils of automation. <http://www.canberratimes.com.au/comment/wealthy-need-to-share-the-spoils-of-automation-20140301-33sp6.html>
- Manyika, J., Chui, M., Bughin, J. & Dobbs, R. (2013). Disruptive technologies: Advances that will transform life, business, and the global economy. McKinsey Global Institute.
- Marcus, G. (2013). Is "Deep Learning" a revolution in artificial intelligence? *The New Yorker*.
- Metzinger, T. (2013). Two principles for robot ethics. In E. Hilgendorf and J. P. Günther (Eds.) *Robotik und Gesetzgebung* (pp. 263-302). Baden-Baden, GER: Nomos.
- Newquist, H. P. (1994). *The brain makers*. Indianapolis, IN: Sams Publishing.
- Norman, D. A. (1986). Reflection on cognition and parallel distributed processing. In J. L. McClelland and D. E. Rumelhart (Eds.) *Parallel distributed processing: Exploration in the microstructure of cognition* (p. 531). Cambridge, MA: MIT Press.
- Padbury, P., Christensen, S., Wilburn, G., Kunz, J. & Cass-Beggs, D. (2014). MetaScan 3: Emerging technologies. *MetaScan 3: Emerging technologies* (p. 45). Cambridge, MA: Policy Horizons Canada.
<http://www.horizons.gc.ca/eng/content/metascan-3-emerging-technologies-0>
- Pinker, S. & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28 (1-2), 73-193. [10.1016/0010-0277\(88\)90032-7](https://doi.org/10.1016/0010-0277(88)90032-7)
- Plewes, T. J. (1990). Labor force data in the next century. *Monthly Labor Review*, 113 (4). Bureau of Labor Statistics, U.S. Department of Labor.
- Print Edition (2014). The future of jobs: The onrushing wave. *The Economist*.
- Robertson, J., Sheppard, T. & Sarnes, S. (2005). Workers compensation claim frequency continues to decline, particularly for smaller claims. *NCCI Research Brief*, 2
- Rumelhart, D. E. (1989). The architecture of mind: A connectionist approach. In M. I. Posner (Ed.) *The architecture of mind: A connectionist approach* (pp. 133-159). Cambridge, MA: MIT Press.
- Rutkin, A. H. (2013). Report suggests nearly half of U.S. jobs are vulnerable to computerization. *MIT Technology Review*.
- Schaal, S., Mohajerian, P. & Ijspeert, A. (2007). Dynamics systems vs. optimal control: A unifying view. *Progress in brain research*, 165 (1), 425-45.
[10.1016/S0079-6123\(06\)65027-9](https://doi.org/10.1016/S0079-6123(06)65027-9)
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3 (03), 417-424. Cambridge, UK: Cambridge University Press. [10.1017/S0140525X00005756](https://doi.org/10.1017/S0140525X00005756)
- Simon, H. A. & Newell, A. (1958). Heuristic problem solving: The next advance in operations research. *Operations Research*, 6 (1), 1-10.
- Stevenson, I. H. & Kording, K. P. (2011). How advances in neural recording affect data analysis. *Nature neuroscience*, 14 (2), 139-42. [10.1038/nn.2731](https://doi.org/10.1038/nn.2731)

- Stunt, V. (2014). Why Google is buying a seemingly crazy collection of companies. *CBC News*.
- Sussillo, D., Abbott, L. F. (2009). Generating coherent patterns of activity from chaotic neural networks. *Neuron*, 63 (4), 544-57. [10.1016/j.neuron.2009.07.018](https://doi.org/10.1016/j.neuron.2009.07.018)
- Thagard, P. (2010). *The brain and the meaning of life*. Princeton, NJ: Princeton University Press.
- (2014). Explanatory identities and conceptual change. *Science & Education*, 23 (7), 1531-1548. [10.1007/s11191-014-9682-1](https://doi.org/10.1007/s11191-014-9682-1)
- Todorov, E. (2008). Optimal control theory. In K. Doya (Ed.) *Bayesian brain: Probabilistic approaches to neural coding* (pp. 269-298). Cambridge, MA: MIT Press.
- (2009). Parallels between sensory and motor information processing. In M. S. Gazzaniga (Ed.) *The cognitive neurosciences* (pp. 613-624). Cambridge, MA: MIT Press.
- Ulam, S. (1958). Tribute to John von Neumann. *Bulletin of the American Mathematical Society*, 64 (3), 5.
- Vinge, V. (1993). The coming technological singularity: How to survive in the post-human era. *Vision 21: Interdisciplinary Science and Engineering in the Era of Cyberspace* (pp. 11-22). Westlake, OH: NASA Conference Publication 10129.

Future Games

A Commentary on Chris Eliasmith

Daniela Hill

In this commentary, the future of artificial minds as it is presented by the target article will be reconstructed. I shall suggest two readings of Eliasmith's claims: one regards them as a thought experiment, the other as a formal argument. While the latter reading is at odds with Eliasmith's own remarks throughout the paper, it is nonetheless useful because it helps to reveal the implicit background assumptions underlying his reasoning. For this reason, I begin by "virtually reconstructing" his claims as an argument—that is, by formalizing his implicit premises and conclusion. This leads to my second claim, namely that more than technological equipment and biologically inspired hardware will be needed to build artificial minds. I then raise the question of whether we will produce *minds* at all, or rather functionally differentiated, fragmented derivatives which might turn out not to be notably relevant for philosophy (e.g., from an ethical perspective). As a potential alternative to artificial minds, I present the notion of postbiotic systems. These two scenarios call for adjustments of ethical theories, as well as some caution in the development of already-existing artificial systems.

Keywords

Artificial minds | Artificial systems ethics | Biological cognition | Mindedness | Postbiotic system

Commentator

[Daniela Hill](#)

daniela.hill@gmx.net

Johannes Gutenberg-Universität
Mainz, Germany

Target Author

[Chris Eliasmith](#)

celiasmith@uwaterloo.ca

University of Waterloo
Waterloo, ON, Canada

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

This commentary has two main aims: First, it aims to reconstruct the major important predictions and claims Eliasmith presents in his target article as well as his reasons for endorsing them. Second, it plays its own version of "future games"—the "argumentation game"—by taking some suggestions presented by Eliasmith maximally seriously and then highlighting problems that might arise as a consequence. Of course, these consequences are of a hypothetical nature. Still, they are theoretically relevant for the question of what will be needed to build full-fledged artificial cognitive agents.

Chris Eliasmith discusses recent technological, theoretical, and empirical progress in re-

search on Artificial Intelligence and robotics. His position is that current theories on cognition, along with highly sophisticated technology and the necessary financial support, will lead to the construction of sophisticated-minded machines within the coming five decades ([Eliasmith this collection](#), p. 2). And also vice versa: artificial minds will inform theories on biological cognition as well. Since these artificial agents are likely to transcend humans' cognitive performance, theoretical (i.e., philosophical and ethical) as well as pragmatic (e.g., legal and cultural laws etc.) consequences have to be considered throughout the process of developing and constructing such machines.

The ideas Eliasmith presents are derived from developments in three areas: technology, theory, and funding; and I will demonstrate the background assumptions underlying these. In this way, I want to demonstrate that if we read Eliasmith as defending a formal argument (rather than a thought experiment), this argument has the form of a *petitio principii*. To illustrate this very clearly, a formal reconstruction of the (not explicitly endorsed, but implicitly assumed) arguments will be conducted. I then argue that even though they are constructed as arguments, and Eliasmith's claims fail, his suggestions provide an insightful contribution to the philosophical debate on artificial systems and the near future of related research. I further want to stress that we should perhaps confine ourselves to talking about less radical alternatives that do not necessarily include the mindedness of artificial agents, but have some element of biological cognition (architecture or software) in them. A number of subordinate questions have to be looked at in order to arrive at a point where a justified statement about the possibility of phenomenologically convincing artificial minds can be made. These considerations include more possibilities than simply the dichotomy of human-like vs. artificial. This is due to our *having* to think about possibilities that lie between or beyond these two extremes, such as fragmented minds and postbiotic systems, since they might soon emerge in the real world. The way in which these will be relevant to philosophy will be largely a question of their psychological make-up—most notably, their ability to suffer.

To start with, the following two sections will present some relevant aspects of the position expressed in the target article. They will summarize, and highlight some of the article's many informative and noteworthy suggestions. I shall also bring in some additional thoughts that I consider important. Afterwards, I will play a kind of future game of my own: I take Eliasmith's predictions very seriously and point at some of the problems that might arise if we were to take his suggestions as arguments. To be fair, [Eliasmith](#) himself says that what he presents are “likely wrong” predictions ([this col-](#)

[lection](#), p. 3). So on a more charitable reading, his claims are not intended to be arguments at all. Yet the attempt to reconstruct them as a formal argument has the advantage of showing that his claims are based on a reasoning that is itself problematic.

2 Are artificial minds just around the corner?

[Eliasmith's](#) perspective on the architecture of minds is a functionalist one ([this collection](#), p. 2, p. 6, pp. 6–7, pp. 9–11, p. 13). The thread running through his paper is his interest in “understanding how the brain functions” and realizing “detailed functional models of the brain” ([ibid.](#), p. 9). The basic idea is that if we construct artificial minds and endow them with certain functions (such as natural language and human-like perceptual abilities), we can examine empirically, in a process comparable to reverse engineering, what it is that constitutes so-called mindedness ([ibid.](#), p. 11). But in their striving to unearth the nature of mindedness, it is not the task of artificial intelligence research or biology to deliver comprehensive and full-fledged theories on biological cognition in general and human cognition in particular. Rather, a very interesting reciprocal relationship between the two parties, in which one learns from the other, is what will propel forward our understanding of biological cognitive systems. In the following I give an overview of the most relevant points that are presented in the target article. They will be divided up into the original sections (technical, theoretical, and empirical).

First, in the technical area and according to Eliasmith, we are fairly far advanced—although there are certain hindrances to successfully implementing theories on this technology. The main obstacle is the size of artificial neuronal systems and, connected to that, their power consumption. Even though neuromorphic chips are being improved steadily, the number of neurons that can be reproduced artificially is still much lower than the number of neurons a human brain has. Thus, the processing of information is significantly slower than in natural cognitive systems ([Eliasmith this collection](#), p.

14). Consequently, what can be realized in the field is still far from the complexity displayed by natural, biological cognition. However, as Eliasmith argues, since we are already in possession of the theoretical groundwork, the main barrier to overcome are technological advances (*ibid.*, p. 9). Throughout the paper, Eliasmith informs the reader that in case we had the technologies needed, artificial minds would immediately be created (*ibid.*, e.g., p. 7, p. 9, p. 11). However, where Eliasmith emphasizes technological barriers, I would like to point out that *theoretical* obstacles exist as well. These mainly revolve around the fact that a system of ethics has to be created *before* we encounter artificial agents. Eliasmith also comments on the consequences for philosophy, arguing that some major positions in the philosophy of mind, such as functionalism, will receive more empirical grounding (*ibid.*, p. 11).

It seems as if the tacit understanding that [Eliasmith](#) has of the function of artificial minds is that they serve as shared research objects of biology and artificial research science in order to gain a better understanding of biological cognition ([this collection](#), p. 9). That is of course only true if indeed the functional architecture of the artificial agent produces convincing behavior, similar to that of biological cognitive systems (humans and animals alike). To illustrate possible problems, one can think of the fact that in research, we learn from animal experiments, even though these animals are quite different from us in many ways. They are, however, similar or at least comparable in one epistemically relevant and specific aspect, i.e., the one that is to be examined, for example in certain aspects of metabolism used to test whether a new drug causes liver failure in humans ([Shanks et al. 2009](#), p. 5). It is the same with artificial agents: they are similar to us in their behavior and thus a worthwhile research object. As such, we could formulate the underlying reasoning as a variant of analytical behaviorism. Analytical behaviorists suppose that intrinsic states of a system are mirrored in certain kinds of behavior. Two systems displaying identical behavior on the outside can be investigated in order to detect whether they do so on the inside

as well ([Graham 2010](#)). This means that we could gain insight on the origin of mental states from a functionally isomorphic system, i.e., an artificially constructed system that is identical in organization and behavior to the natural system copied.

Last, since it seems that it will be possible in the future, given the required hardware, to design artificial agents according to our needs, it does not appear far-fetched to assume that the quality of human life might consequently be improved to a great extent ([Eliasmith this collection](#), p. 11). This requires, however, that we make up our own minds about how to interact with such agents, which rights to grant and which to deny them. And also the opposite case may not be disregarded: it is imaginable that the artificial agents will at some point turn the tables and be the ones to decide on *our* rights (cf. [Metzinger 2012](#)). In highlighting aspects from different areas to be considered, Eliasmith reminds us of the possibilities that lie ahead of us, but also of the challenges that might show up and have to be faced. I want to suggest that we also take into consideration alternative outcomes that are not minds in the biological sense, but rather derivatives of minds. I will therefore put the notion of postbiotic systems into play as a way of escaping the dichotomy “human-like” vs. “artificial” ([Metzinger 2013](#)). The philosophical point here is that the conceptual distinction between “natural” and “artificial” may well turn out to be non-exhaustive and non-exclusive: there might well, as Metzinger points out, be future systems that are neither artificial nor biological. By no means do I intend to argue against the use of scientific models, since they are what good research needs. Rather, I wish to draw attention to the possible emergence of intermediate systems, rather than only the extremes (i.e., human-like vs. artificial agents), or classes of systems that go beyond our traditional distinctions, but which nevertheless count as “minded”. As mentioned above, this is due to these intermediate or postbiotic systems being possible much earlier—probably preceding full-blown minded agents.

I will end this section by drawing attention to some of the author’s thoughts on the crucial elements of artificially-minded systems. According

to Eliasmith, three types of skills are vital in building artificial minds: cognitive, perceptual, and motor skills have to be combined to create a certain behavior of the minded artificial agent. This behavior will then serve as the basis for us humans to judge whether we perceive the artificial agent as “convincing” or not (Eliasmith this collection, p. 9). Unfortunately, no closer specification of what it is to be “convincing” is given in the target article. No theoretical demarcation criterion is offered. What we can say with great certainty, however, is that in the end our subjective *perception* of the artificial agents will be the decisive criterion. One could speculate on whether it is merely an impression, or even an illusion, that leads us to concluding that we are facing a *minded* agent. According to Eliasmith, any system that produces a robust social hallucination in human observers will count as possessing a mind.

3 Playing the “argumentation game”

In the following I will play the “argumentation game” and for a moment assume that what Eliasmith presents us with actually is argumentation. The goal of this section is not to claim that Eliasmith really *argues* for the emergence of artificial minds in the classical way. Rather, I wish to highlight that possibly more than technological equipment and biologically inspired hardware need to be taken into account before research can present us with a mind, as outlined by Eliasmith. If we deconstruct his line of reasoning and virtually formalize the *argument*, we don’t find valid argumentation but rather a set of highly educated—and certainly informative—claims about the future, which doubtlessly help us prepare for a future not too far ahead of us. I will utilize the terms “argumentation”, “argument”, “premise”, and “conclusion” in the following, but it should always be remembered that these terms are only “virtually” or hypothetically. So let us see how Eliasmith proceeds:

If we play the argumentation game, a first result is that Eliasmith’s virtual argument becomes problematic at the moment he starts elaborating on theoretical developments that have been made and that will propel forward the development of “brain-like models” (this collection,

p. 6). From the perspective of an incautious reader, the entire section “Theoretical developments” could be seen as resulting in a claim that can be traced back to a *petitio principii*. This means that the conclusion drawn at the end of the argumentative line is identical with at least one of the implicit premises. The implicit argumentation is made up of three relevant parts and unfolds as follows: first, building brain-like models is not only a matter of the available technological equipment (*ibid.*, first paragraph; cf. premise 1). Instead, if we face a convincing artificially-minded agent, it is characterized by both sophisticated technological equipment and by our discovery of principles of how the brain functions, such as learning or motor control (*ibid.*; cf. premise 2). And so, in conclusion, it follows that if biological understanding and technological equipment come together, we will be able to build brain-like models and implement them in highly sophisticated cognitive agents (*ibid.*).

The incautious reader would now have to believe that Eliasmith is confusing necessary and sufficient conditions. Let us look at this assumed argument in some more detail. Formulated as a complete argument we would get: “If it is not the case that technological equipment *alone* leads to the building of brain-like models for artificial cognitive agents, but we face a good artificial minded agent which is endowed with certain technology as well as biologically inspired hardware, we have to conclude that this certain technology and biologically inspired hardware are not only necessary, but also sufficient for building brain-like models for artificial cognitive agents.”

The formal expression of this argument would be the following:

T: We have developed sophisticated technological equipment.

B: We have developed biologically-inspired hardware.

M: We can build brain-like models which can be implemented in artificial cognitive agents.

$$\begin{array}{l} \neg(T \rightarrow M) \\ M \rightarrow (T \ \& \ B) \\ \hline (T \ \& \ B) \rightarrow M \end{array}$$

As is obvious from how the argument is constructed, it is invalid. So, what we can say at this point is that the combination of both technical features and biologically-inspired neuromorphic hardware very likely does get us some way, but we might have to consider which elements are missing so that we really end up building what will be perceived as minds. I shall propose some possibilities in the following section. The author even supposes that we will be able to build artificial agents ready to rival humans in cognitive ability (Eliasmith [this collection](#), p. 9). I am convinced that it is not cognitive artificial agents that will be the crucial hurdle, but rather their mindedness. I am also convinced that the huge amount of money spent on certain research projects will most likely result in improved models of the brain, as suggested by Eliasmith ([ibid.](#), p. 8), but it is not obvious to me how investing a vast amount of money necessarily results in relevant findings. It is also possible that no real progress will be made. Stating the opposite, which Eliasmith does not, resembles a claim based on expertise as bulletproof evidence. Sure enough, monetary sources are needed to make progress, but they are no *guarantee*. So possibly technology, biological theories on the brain's functioning, and money, essentially, might not lead to sophisticated cognitive agents being built ([ibid.](#)). The point is not that we should not invest money unless a positive outcome is guaranteed. Rather, we need a theoretical criterion for mindedness that is philosophically convincing—and not only robust, but epistemically unjustified social hallucinations. This theoretical criterion is what we lack.

4 What could artificial minds be?

In this section, I intend to sketch some important issues and questions for the future debate on artificial minds. I shall examine whether predictions on the concept of *artificial minds* can be made at the present state of the debate and based on the empirical data we currently have. This involves knowledge about what a mind is, and knowledge about how an *artificial* mind is characterized. In reconstructing Eliasmith's un-

derstanding of what a mind is, we may find the following statement informative: he relies on behavioral, theoretical, and similarity-based methods ([this collection](#), p. 3). The possible problem with this approach is that the characterization of the methods is very limited. To point to some relevant questions: what is the behavior of a mind? What about the fact that *mind* is not even close to being well understood theoretically? How do similarity-based methods avoid drawing problematic conclusions from analogies (cf. Wild 2012)? Importantly, at this point we are only talking about natural, biologically-grounded minds. Answers as to what an *artificial* mind is supposed to be might exceed the concept of mind in ways we are unable to tell at the present moment.

Let us see how Eliasmith characterizes artificial *minds*. One can see this as a judgment based on the similarity of behavior originating from two types of agents: humans and artificial. Functions need to be developed that are necessary for building an artificial mind. These functions lead to a certain kind of behavior. This behavior is achieved by perceptive, motor, and cognitive skills, which are needed to make the behavior seem human-like. Thus, the functions implemented on sophisticated kinds of technology will, in the end, lead to human-like behavior (Eliasmith [this collection](#), p. 9). The reason why the argumentative step from cognition, perception, and motor skills to mindedness can be made is the underlying assumption that the behavior resulting from these three types of skills is *convincing* behavior in our eyes (Eliasmith [this collection](#), p. 10). Similarity judgments, so Eliasmith argues, might appear “hand-wavy”. Still, he uses them to reduce the complexity that mindedness brings with it ([ibid.](#), pp. 5–6), and he certainly succeeds in drawing attention to a whole range of important issues. However, it could well be that the reduction to human-like behavior as the benchmark for assessing mindedness is too simple. After all, analytical behaviorism today counts as a failed philosophical research program. There could be much more to mindedness than behavior. We just do not know what this is yet. As a possible candidate we might consider the previously

mentioned psychological make-up of artificial agents, such as their being endowed with internal states like ours. One might think of robust first-person perspectives, but also about emotions like pain, disappointment, happiness, fear, and the ability to react to these. Other options include interoceptive awareness or the ability to interact socially—and much more.

5 What should we brace ourselves for?

Given the complexity of mindedness and our very limited understanding of what constitutes it, what else can we talk about? We could consider further possibilities of artificial systems that might arise, thereby enlarging the set of constraints that has to be satisfied. Some of them seem much more likely than artificial minds, and they might precede minds chronologically. I would like to focus on the idea of *fragmented minds* on the one hand and of *postbiotic systems* on the other, as two versions of artificial systems. An artificially-constructed fragmented mind is characterized by only partial satisfaction of the constraints fulfilled by a human mind. It could thus, very much like autistic persons with savant syndrome (i.e., more than average competence in a certain domain, e.g., language learning or music), and possess only some of our cognitive functions, but be strikingly better at them than normal humans are and ever could be, given their biological endowment.¹ Postbiotic minds, on the other hand, could satisfy additional constraints that are not yet apparent presently. I will conclude with some reflections on the new kind of ethics that will have to be created in order to approach new kinds of cognitive agents. As pointed out above, I assume that cognitive agents will be possible much earlier than truly minded agents. Learning, remembering, and other cognitive functions can already be recreated in artificial systems like *Spaun*. Still, human cognition is very versatile and complex. A fully minded agent, in contrast to a merely cognitive agent, might also be able to experience herself as a cognitive agent.

Therefore, I propose that cognitive systems could be created that do not yet qualify as a copy of our cognitive facilities, but which cover only parts of our cognitive setup. I call these *fragmented minds*. Importantly, the word *minds* does not refer here to the artificiality of the system at all. There are human beings with fragmented minds, too, such as babies, who do not yet display the cognitive abilities we ascribe to adult humans in general, or the aforementioned autistic humans with savant syndrome. Fragmented minds are contrasted with what we experience as normal human minds. *Fragmented* means that the created system possesses only part of the abilities that our mind displays. The term *mind* delineates the—historically contingent—point of reference that is human beings. How are fragmented minds further characterized? Eliasmith himself gives us an example: we could design a robot (an artificial mind) that gains fulfillment from serving humans ([this collection](#), p. 11). This would only be possible if aspects of our own minds were not part of the mental landscape of this robot. We could roughly formulate such an aspect, such as the will to design one's own life. Folk psychology would most likely regard this robot as lacking a free will, which is in conformity with the idea of slavery that Eliasmith acknowledges (*ibid.*). So a fragmented mind is an artificial system that possesses part of a biological cognitive system's abilities instead of the rich landscape most higher animals (e.g., some fishes and birds, certainly mammals), as well as humans, display.

Related to the aspect of fragmented minds is the idea that we could refrain from creating minds that might cause us a lot of moral and practical trouble, and instead focus on building sophisticated robots designed to carry out specific kinds of tasks. Why do we need to create artificial *minds*? What is the additional value gained? If these robots are not mindful, we will circumvent the vast majority of conceptual and ethical problems, such as legal questions (What is their legal status compared to ours?) or ethical considerations (If I am not sure whether an artificial agent can perceive pain, how should I treat it in order to not cause harm?). In which case, they

¹ In that case, the variable **B** from above (biologically inspired hardware) would not be a necessary condition for finding out more about mindedness.

would only be more capable technology than what we know at present, and most likely be of no major concern for the philosophy of mind. However, if they *are* mindful, we doubtlessly have to think about new ways of approaching them ethically.

Also ethically relevant are intermediate systems, systems that are not clearly either natural or artificial. These systems have been called *postbiotic systems* (Metzinger 2012, p. 268). What characterizes postbiotic systems is the fact that they are made up of both natural and artificial parts, thus belonging to neither of the exhaustive categories “natural” or “artificial”. In that way a natural system, e.g., an animal, could be controlled by artificially-constructed hardware (as in hybrid bio-robotics); or, in the opposite case, artificial hardware could be equipped with biologically-inspired software, which works in very much the same way as neuronal computation (Metzinger 2012, pp. 268–270; Metzinger 2013, p. 4). Perhaps Eliasmith’s own brain-like model *Spaun* is a postbiotic system in this sense, too. In what way would these systems become ethically relevant? Although the postbiotic systems in existence today do not have the ability to subjectively experience themselves and the world around them, they might have it in the future. In being able to subjectively experience their surroundings, they are probably also able to experience the state of suffering (Metzinger 2013, p. 4). Everything that is able to consciously experience a frustration of preferences as a frustration of its *own* preferences automatically becomes an object of ethical consideration, according to this principle. For such cases, we have to think of ethical guidelines *before* we are confronted with a suffering postbiotic mind, which could be much earlier than we expect. Before thinking about how to implement something as complex and unpredictable as an artificial *mind*, one should consider what one does *not* want to generate. This could, for example, be the ability to suffer, the inability to judge and act according to ethical premises, or the possibility of developing itself further in a way that is not controllable by and potentially dangerous for humans.

6 Conclusion

In this commentary, I have played the “argumentation game” as my own version of Eliasmith’s “future game”. The intention behind this was to demonstrate that we very likely need more than sophisticated technology and biologically-inspired hardware to build brain-like models ready to be applied in artificial cognitive agents. As such, I playfully took Eliasmith’s considerations on the future of artificial minds as arguments, and demonstrated that they would result in a *petitio principii*. In so doing, I highlighted that necessary conditions do not have to be sufficient as well. While this is common philosophical currency, it is instructive to spell this out in the case of artificial agents. So in the present case, what constitutes artificial cognitive systems and what is needed to gain a deeper understanding of how the mind works might include more factors than the two crucial ones Eliasmith outlines, namely biological understanding and its implementation in highly-sophisticated technology. I proposed some possibilities that might turn out to be informative for future considerations on what constitutes an artificial mind. In particular, I mentioned experiential aspects, such as the perception of emotions and reactions to them, as well as internal perceptions like interoceptive awareness. In general, this means that we need theoretical criteria that are convincing for philosophy in order to overcome referring to robust yet convincing social hallucinations. Further, to illustrate that the distinction between natural and artificial systems might not be exhaustive, I pointed to the notions of fragmented minds and postbiotic systems as possible developments for the nearer future. They have to be considered, in particular with respect to their ethical implications, before they are developed and implemented in practice.

Even though we lack a more fine-grained, deeper understanding of what constitutes minds, Eliasmith shows us that it is worth thinking about what we already *do* have at hand for constructing artificially-minded systems. He demonstrates vividly that two factors—technology and biology—are of major import-

ance on the route to artificially-cognitive, if not minded, agents. And he brings into discussion a number of far-reaching consequences that will apply in case we do succeed in building artificial minds within the next five decades. These will inform the development of these artificial systems as well as philosophical debate, both on an ethical, as well as theoretical level. In this way, Eliasmith's contribution has to be regarded as significant in terms of preparing us for the decades to come.

Acknowledgements

First and foremost, I am grateful to Thomas Metzinger and Jennifer M. Windt for letting me be part of this project, thus providing me a unique opportunity to gain valuable experience. Further special thanks go to the two anonymous reviewers, as well as the editorial reviewers for their insightful comments on earlier versions of this paper. Lastly, I wish to express my gratitude to Anne-Kathrin Koch for sharing her expertise with me.

References

- Eliasmith, C. (2015). On the eve of artificial minds. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Graham, G. (Ed.) (2010). Behaviorism. *The Stanford Encyclopedia of Philosophy*.
<http://plato.stanford.edu/entries/behaviorism/>
- Metzinger, T. (2012). *Der Ego-Tunnel: Eine neue Philosophie des Selbst: Von der Hirnforschung zur Bewusstseinsforschung*. Berlin, GER: Bloomsbury.
- (2013). Two principles for robot ethics. In E. Hilgendorf & J.-P. Günther (Eds.) *Robotik und Gesetzgebung* (pp. 263-302). Baden-Baden, GER: Nomos.
- Shanks, N., Greek, R. & Greek, J. (2009). Are animal models predictive for humans? *Philosophy, Ethics, and Humanities in Medicine*, 4 (2), 1-20.
[10.1186/1747-5341-4-2](https://doi.org/10.1186/1747-5341-4-2)
- Wild, M. (2012). *Fische. Kognition, Bewusstsein und Schmerz: Beiträge zur Ethik und Biotechnologie*. Bern, CH: Bundesamt für Bauten und Logistik BBL.

Mind Games

A Reply to Daniela Hill

Chris Eliasmith

In her discussion of my original article, Hill reconstructs an argument I may have been making, argues that the distinction between natural and artificial minds is not exclusive, and suggests that my reliance on behaviour as a determiner of “mindedness” is a dangerous slip back to philosophical behaviourism. In reply, I note that the logical fallacy of which I’m accused (circular reasoning) is not the one present in the reconstruction of my argument (besides the point), and offer a non-fallacious reconstruction. More importantly, I note that logical analysis does not seem appropriate for the discussion in the target article. I then agree that natural and artificial minds do not make up two discrete categories for mindedness. Finally, I note that my research program belies any behaviourist motivations, and reiterate that even though behaviour is typically important for identifying minds, I do not suggest that it is a substitute for theory. However, the target article is not about such theory, but about the near-term likelihood of sophisticated artificial minds

Keywords

Artificial minds | Behaviourism | Logical analysis | Minds

Author

[Chris Eliasmith](#)

celiasmith@uwaterloo.ca

University of Waterloo
Waterloo, ON, Canada

Commentator

[Daniela Hill](#)

dhill@students.uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

I think Hill is right to wonder aloud about my methodology in the target article. After all, I just ignored the hard philosophical issue of saying what minds really are. I pretended (somewhat self-consciously) that we all know what minds are, and so that we will simply be able to tell when someone has created one, if they ever do. But, I did that for a reason. The reason was this: I did not want to get lost in the minutiae of metaphysics when my focus was on a technological revolution—one with significant philosophical consequences (which is also not to say I don’t like such minutiae in their proper time and place).

2 A failure of logic

However, Hill was also not especially taken by the reasons I provided for expecting such developments either. Hill’s suggestion is that the best reasonable argument you could construct from my original considerations was fallacious. Though, [Hill](#) is quick to point out that I didn’t take myself to be constructing an argument: “... not to claim that Eliasmith really argues for the emergence of artificial minds” ([this collection](#), p. 4).

Nevertheless, her analysis is that what I have provided is best understood as a *petitio principii* (aka circular argument): “this means that the conclusion drawn at the end of the ar-

gumentative line is identical with at least one of the implicit premises” (Hill [this collection](#), p. 4). Unfortunately, the technical analysis offered (p. 5) is a *non sequitur* (i.e., there is no logical connection between the premises and conclusion). Regardless, one fallacy is as embarrassing as the other.

However, I’d like to suggest that if we wanted to recast the original paper as a logical argument, then a simple *modus ponens* will do: if we have a good theory and the technological innovations necessary to implement the theory, then we can build a minded agent. We have good (and improving) theory and will have the proper technological innovations (in the next 50 years), therefore we will be able to build a minded agent (in the next 50 years). Indeed, most of the paper is arguing for the plausibility of these premises.

More to the point, however, I think that we can take this as an object lesson for when logical inference is really just the wrong kind of analysis of a paper. Instead of trying to provide a logical argument from which the conclusion necessarily follows from the premises, I am providing series of considerations that I believe make the conclusion likely given both the current state of affairs, and expected changes. In short, I think of the original paper as providing something more like a series of inferences to the best explanation: all of which are, technically, fallacious; and all of which are directed at establishing premises.

3 Back to minds

Despite disagreeing with the analysis of the logical structure of the paper, I do appreciate the emphasis that Hill has placed on philosophical and ethical aspects of our attempts to construct minds. In the original article, I only very briefly touch on those issues. However, I would be quick to point out that I do not think, and never intended to suggest, that the distinction between “natural” and “artificial minds” was an absolute, “exhaustive,” or “exclusive” one (Hill [this collection](#), p. 3). Like most interesting and complex features, possession of ‘mindedness’ no doubt comes in degrees. In fact, I think that our

attempts to construct artificial minds will provide a much better sense of the dimensions along which such a continuum is best defined.

Finally, I must admit that I find it somewhat alarming that I’m being characterized as a behaviourist in Hill’s article—*that* has definitely never happened before: “Let us see how Elia-smith characterizes artificial minds. One can see this as a judgment based on the similarity of behaviour originating from two types of agents: humans and artificial” ([this collection](#), p. 5). Hence, I was espousing “analytical behaviorism... a failed philosophical research program” (Hill [this collection](#), p. 6). Indeed, I, like all behavioural scientists, believe that behaviour is one important metric for characterizing the systems of interest. However, the reason I focus on internal mechanisms in my own research – all the way down to the neural – is that I believe those mechanisms give us critical additional constraints for identifying the right class of algorithms that give rise to behaviours. Consequently, I wholeheartedly agree with Hill that “There could be much more to mindedness than behaviour” ([this collection](#), p. 6). So, for the record, I believe that our best theories for how to build minds are going to be highly informed by low-level mechanisms.

That being said, I also think that most people’s judgments of whether or not something counts as being minded is going to come down largely to their being convinced of the naturalness, or “cognitiveness” of the behaviour that is exhibited by agents we construct. Notice that there is a difference between a claim of how people will judge mindedness, and a claim about theories of mindedness or how we ought to best achieve that judgment. Turing was, after all, onto *something* with his test.

4 Conclusion

I noted in the original article (Eliasmith [this collection](#)) that I was attempting to avoid becoming mired in tangential debates regarding what it is to have a mind by simplifying the criteria for mindedness (for the purposes of that article). Exactly the kinds of debates I was attempting to avoid are raised in Hill’s comment-

ary. For example, I don't think we know if there is a clean contrast between a "fully minded agent" and a "merely cognitive agent" ([Hill this collection](#), p. 6). Perhaps there is, and perhaps it is that a fully minded agent can "experience herself as a cognitive agent", ([Hill this collection](#), p. 6) but perhaps not. This does not strike me as a decidable question at present.

So, perhaps my unwillingness to venture into the murky waters of necessary and sufficient conditions for having a mind came off as making me look like a behaviourist. But in truth, my purpose was rather to focus on providing a variety of evidence that I think suggests that artificial minds are not as far away as some have assumed. There is, I believe, a historically unique confluence of theory, technology, and capital happening as we speak.

References

- Eliasmith, C. (2015). On the eve of artificial minds. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Hill, D. (2015). Future games - A commentary on Chris Eliasmith. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.

Can We Be Epigenetically Proactive?

Kathinka Evers

The human brain is an essentially evaluative organ endowed with reward systems engaged in learning and memory as well as in higher evaluative tendencies. Our innate species-specific, neuronally-based identity disposes us to develop universal evaluative tendencies, such as self-interest, control-orientation, dissociation, selective sympathy, empathy, and xenophobia. The combination of these tendencies may place us in a predicament. Our neuronal identity makes us social, but also individualistic and self-projective, with an emotional and intellectual engagement that is far more narrowly focused in space and time than the effects of our actions. However, synaptic epigenesis theories of cultural and social imprinting on our brain architecture suggest that there is a possibility of culturally influencing these predispositions. In an analysis of epigenesis by selective stabilisation of synapses, I discuss the relationships between genotype and brain phenotype: the paradox of non-linear evolution between genome and brain complexity; the selection of cultural circuits in the brain during development; and the genesis and epigenetic transmission of cultural imprints. I proceed to discuss the combinatorial explosion of brain representations, and the channelling of behaviour through “epigenetic rules” and top-down control of decision-making. In neurobiological terms, these “rules” are viewed as acquired patterns of connections (scaffoldings), hypothetically stored in frontal cortex long-term memory, which frame the genesis of novel representations and regulate decision-making in a top-down manner. Against that background I propose the possibility of being epigenetically proactive, and adapting our social structures, in both the short and the long term, to benefit, influence, and constructively interact with the ever-developing neuronal architecture of our brains.

Keywords

Cultural circuits | Empathetic xenophobia | Epigenetic proaction | Epigenetic rules | Neuroethics | Precaution | Responsibility | Selective sympathy | Species-specific identity | Synaptic epigenesis

1 Introduction

Contemporary neuroscience no longer views the brain as an input-output processing device but as an autonomously active, self-referential, and selectional system operating in a projective style, which is in constant social interaction and in which values are incorporated as necessary constraints. The idea that evolution by natural selection has given rise to an essentially evaluative cerebral architecture raises the question whether, in the human species, such neurobiologically-based predispositions have further developed the means to generate novel specific values on higher cognitive levels. The concept of “value” would then play a central role as something that is taken into account in decision-

making and that influences a choice, selection, or decision, that can occur on many levels—non-conscious as well as conscious—as a basic biological function or as a feature of advanced moral reasoning. But, if we are born evaluators, to what extent can these predispositions with which we are all born be culturally controlled?

In this article, I suggest that our innate species-specific neurally based identity disposes us to develop universal evaluative tendencies, such as self-interest, control-orientation, dissociation, selective sympathy, empathy, and xenophobia. The combination of these tendencies may place us in a practical and moral predicament. Our neuronal identity as persons makes

Author

Kathinka Evers

kathinka.evers@crb.uu.se

Uppsala Universitet

Uppsala, Sweden

Commentator

Stephan Schleim

s.schleim@rug.nl

Rijksuniversiteit Groningen

Groningen, Netherlands

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität

Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University

Melbourne, Australia

us social, but also individualistic and self-projective, with an emotional and intellectual engagement that is far more narrowly focused in space and time than the effects of our actions.

However, the neuronal organisation of our adult brain develops in the course of a fifteen year-long period following birth, during which, and, to a lesser extent, after which it is subject to cultural influence, both on the individual level and, at the social group level, across generations (Lagercrantz 2005; Lagercrantz et al. 2010; Collin & van den Heuvel 2013). Synaptic epigenesis theories of cultural and social imprinting on our brain architecture (which differ from less discriminative epigenetic modifications of nuclear chromatin) (Changeux 1985; Kitayama & Uskul 2011) suggest that there is an interesting possibility, which, in my opinion, has hitherto been underestimated. That is, we could potentially be *epigenetically proactive* (Evers 2009) and adapt our social structures, in both the short and the long term, to benefit, influence, and constructively interact with the ever-developing neuronal architecture of our brains.

2 The social individualist

2.1 An egocentric evaluator

The human brain is intrinsically active: it produces electrical and chemical activity both in response to external stimuli and, spontaneously, independently of them. The brain is an autonomously-active motivated neuronal system, genetically equipped with a predisposition to explore the world and to classify what it finds there (Changeux 1985, 2004). On-going spontaneous activity is present throughout the nervous system. In the embryo, spontaneous movements (Narayanan & Hamburger 1971) and waves of endogenous retinal activity (Galli & Maffei 1988; Goodman & Shatz 1993) are thought to play an important role in the epigenesis of neural networks through synapse selection (see below). On-going spontaneous activity is also present in the adult brain, where it is responsible for the highly variable patterns of the electroencephalogram (EEG; Berger 1929; Raichle et al. 2001). Thalamocor-

tical networks generate a variety of oscillations, whose rhythms change across the sleep-wake cycle (Llinas & Paré 1991). Optical imaging methods in anesthetized animals also reveal fast spontaneous states of neuronal activity that, far from being random, exhibit patterns that resemble those evoked by external stimuli. In parallel, functional neuroimaging studies in humans have shown a globally-elevated brain metabolism at rest, with localized patterns suggesting that particular cortical regions are maintained in a high, although variable, state of activity referred to as “default mode” by Raichle et al. (2001).

Hypotheses of knowledge acquisition posit that patterns of spontaneous activity, referred to as “pre-representations”, arise in the brain and are selected by reward signals as “representations” confirmed by both external experience and internal processes of evaluation within a conscious neuronal workspace (Dehaene & Changeux 2011). Such “models of the world” are stabilised through “cognitive games” by analogy with Wittgenstein “language games”, as permanent features of the developing cognitive apparatus, according to a process referred to as “mental Darwinism” (Changeux 2004).

Anticipation of reward signals introduces a delay between the elaboration of tacit plans of action and actual interaction with the world performed by the organism, which presupposes a distinction of temporal states: awareness of the present, remembrance of the past, and anticipation of the future (Barto & Sutton 1982; Schultz et al. 1997; Dehaene & Changeux 2000; Schultz 2006). Without any capacity to evaluate stimuli, the brain could neither learn nor remember: it has to prefer some stimuli to others in order to learn. This classical idea in learning theory has been expressed in neuronal terms by Dehaene & Changeux (1991), and by Edelman in his accounts of primary consciousness (Edelman 1992). In these accounts, learning is a change in actual behaviour, or the storage of a trace subsequently unveiled (Dudai 1989, 2002) through brain categorizations of stimuli. These are given in terms of positive or negative values, understood as something that is taken into account in decision-making and that influences a choice, selection or decision, which can occur on

many levels. Through its intense and spontaneous activity, the brain has also been described as a narrative organ, spinning its own neuronal tale (Evers 2009). The narrations will vary greatly between individuals, but each will be self-projective.

The natural egocentricity or individualism of the human brain appears quite pronounced. In its projection of autonomously-produced images, the brain refers all experiences to itself, that is, to its own individual perspective. This self-projection is a biological predisposition that humans possess innately and that is closely connected to our predisposition for developing self-awareness, which Edelman suggests is a necessary condition for developing higher-order consciousness (Edelman 1992; Denton 2006; see also Tulving 1983). The existence of a self-projecting systems monitoring internal processes in the brain was suggested by an early Positron Emission Tomography (PET) study of self-generated actions showing hemodynamic activity in the posterior cingulate cortex (Blakemore et al. 1998). This observation was confirmed and extended by magneto-encephalography following synchronization in the gamma range (55–100 Hz), thus defining a major network of the brain: the paralimbic interaction between the medial prefrontal/anterior cingulate and medial parietal/posterior cingulate cortices and subcortical regions (Lou et al. 2004; rev. Changeux & Lou 2011). Damasio (1999) distinguished a “core consciousness” (core self) from an “extended consciousness” (extended self) that we consider as analogous to the “minimal self” and “extended self” of Gallagher (2000). Minimal self-awareness is prereflexive, immediate and normally reliable, while still involving a sense of ownership of experience (Gallagher 2000). The “extended self” is a coherent self that persists across time and requires a system that can retrieve long-term memories of personal experiences—namely, episodic memory (Gardiner 2001). Consequently, episodic memory retrieval becomes an indispensable component of the more complex forms of self-awareness and consciousness (Tulving 1983).

In the course of growing up, the infant develops the capacity to focus its attention; it

learns to distinguish between and recognise objects in its environment, such as faces, and becomes aware of itself as standing in various relations to these objects. Conscious processing develops into auto-distinction (when “this-here” is distinguished from “that-there”). When further developed, the individual becomes aware of itself as a subject of experience and ascribes mental states to itself: auto-distinction evolves into self-awareness (when “this-here” becomes “I”) usually at around one and a half years of age (Lagercrantz 2005), and possibly even earlier (Falck-Ytter et al. 2006; see also Rochat 2001). From the age of six to twelve months, the child typically sees a “sociable playmate” in the mirror’s reflection. Self-admiring and embarrassment usually begin at twelve months, and at fourteen to twenty months most children demonstrate avoidance behaviours. Finally, at eighteen months 50% of children recognize the reflection in the mirror as their own and by twenty to twenty-four months this rises to 65%—this is revealed, for instance, by them trying to evince marks on their own nose, taking advantage, in all these instances, of their episodic memory abilities (see Tulving 1983).

An evolved survival function that adds an evaluative element to our brain’s self-projective mode of operation is self-interest, expressed as a desire to survive, to be well-fed, safe, to reproduce, and so on. This is not a defining characteristic, for there are exceptions, for example subjects who have a very poorly developed self-interest (Damasio 1994; Damasio & Carvalho 2013). Nor is it necessarily rational, since biological evolution is circumstantial. There is an abundant literature on the phenomenologically rich concept of self-interest in philosophy and ethics, in terms e.g., of enlightenment, egoism, capacity for altruism, etc. Such issues are relevant and interesting but beyond the scope of this discussion. In the present context, self-interest is understood in a minimalistic sense, as an evolved survival function that adds an evaluative element to our brain’s self-projective mode of operation.

Self-interest is also a source of the urge to control the immediate environment, and of the need for familiarity, security, and preference for

the known. The subjective experience of some level of control and the security that this provides is in fact a necessary condition for the individual to develop in a healthy manner and to consolidate an integrated sense of self (Le-doux 1998). When the external circumstances become severely disturbing, we feel increasingly threatened and have a defence mechanism that is eventually activated: *dissociation*, here understood as a process whereby information—incoming, stored, or outgoing—is actively prevented from integration with its usual or expected associations.

The human being is, in this sense, a “dissociative animal”: we spend a considerable amount of intellectual and emotional energy on distancing ourselves from a wide range of things that we consciously or non-consciously fear or dislike (Evers 2009). When an experience is too painful to accept, we sometimes deliberately do not accept it; instead of integrating it into our ordinary system of associations, we push it away from us, and prevent it from being integrated into our consciousness. Pushed to an extreme, this tendency may become pathological, e.g., in the development of Dissociative Identity Disorder (cf. DSM-IV), but as a non-pathological process it is an important adaptive function, and a valuable evolutionary asset allowing us to survive events that we would otherwise be unable to endure (Putnam 1989; Evers 2001).

So far, I have described the brain as an autonomously active, self-projective, and selectional neural system with innate evaluative tendencies, e.g., self-interest, control-orientation, and dissociation. These cerebral features characterize the individual, but they are also reflected in the social relationships proper to the human species.

2.2 Selective sympathy & empathetic xenophobia

In social animals, self-interest is a source of interest in others. In the case of humans, this social interest focuses primarily on those to whom the self can relate and with whom it identifies, such as the next of kin, the clan, the community, etc. The human brain conjugates op-

posite tendencies: first, embodied in the human subject, it is engaged in highly individualistic and self-projective actions, such as the search for water or food. But it also mediates co-operative social relationships: the “I” is extended to endorse the group, as a “we”, and distinctions are drawn between “us” and “them” (Ricoeur 1992; Changeux & Ricoeur 2000). Sympathy and aid is typically extended to others in proportion to their closeness to us in terms of biology, e.g., face recognition (Michel et al. 2006; Hills & Lewis 2006), racial out-group versus in-group distinctions (Hart et al. 2000; Phelps et al. 2003), culture, ideology, etc.

Imagining an action or actually performing that action both have similar neural circuits (which include the premotor cortex, supplementary motor area, cerebellum, parietal cortex, and basal ganglia) to those activated when one observes, imitates, or imagines actions performed by other individuals (Jeannerod 2006; Decety 2012). The model mechanism suggested is that actions are coded in terms of perceivable effects (Hommel et al. 2001). Performing a movement leaves a memory of the association between the motor pattern by which it was generated and the sensory effects that it produces. Such stored associations can then be used to retrieve a movement by anticipating its effects. This perception-action coupling mechanism, which includes active sensing and motor-sensory loops (Gordon & Ahissar 2012) and to which may be added the motor theory of language (Liberman & Mattingly 1985), offers a mechanism for intersubjective communication and social understanding by creating functional links between first-person and third-person information (Decety & Sommerville 2003; Jackson & Decety 2004).

Functional Magnetic Resonance Imaging and magneto-encephalography among other methods have led to the demonstration that when children or adults watch other subjects in pain, the neural circuits mobilized by the processing of first-hand experience of pain are activated in the observer (Singer et al. 2004; Cheng et al. 2008). This sharing allows mapping of the perceived affective cues of others onto the behaviours and experiences of the self-oriented

response. Decety (2012) argues that, depending on the extent of the overlap in the pain matrix, and complex interactions with personal dispositions, motivation, contextual information, and self-regulation, this can lead to personal distress (i.e., self-centred motivation) or to empathic concern (i.e., an other-oriented response). This basic somatic sensorimotor resonance plays a critical role in the recognition and sharing of others' affective states.

There is an important neural distinction between apprehending and caring that makes it possible to understand the affective state of another without feeling engaged in it. Studies in the neurobiology of empathy (here understood as the ability to apprehend the mental states of others), and sympathy (the ability to care about others) suggest that these abilities involve complex cognitive functions with large individual and contextual variations that depend on both biological and socio-cultural factors (Jackson & Decety 2004; Singer et al. 2004; Singer et al. 2006; Iacoboni et al. 2005; Jackson et al. 2006; Lawrence et al. 2006; Parr & Waller 2006; Engen & Singer 2013). Such results are important, because appreciating the brain's role in apprehending and responding to the affective states of others can help us understand people who exhibit social cognitive disorders and are deficient in experiencing socially relevant emotions such as sympathy, shame, or guilt.

However, even in supposedly healthy human brains the capacity for other-oriented responses, such as sympathy, is pronouncedly selective and limited by spontaneous aggressive tendencies (Panksepp 1998; Lorenz 1963). When sympathy and mutual aid is extended within a group, they are also (de facto) withheld from those that do not belong to this group. In other words, interest in others is ordinarily expressed positively or negatively towards specific groups—but very rarely are attitudes extended to universal coverage, for example as attitudes towards the entire human species, or towards all sentient beings.

Understanding does not entail compassion, but is frequently combined with emotional dissociation from “the other”. We can easily understand, say, that a child in a distant country

probably reacts to hunger or pain in a way that is similar to how children in our own country react to hunger or pain, but that does not mean that we care about those children in equal or even comparable measures. Indeed, if understanding entailed sympathy, the world would be a far more pleasant dwelling place for many of its inhabitants. By nature, we are “empathetic xenophobes” (Evers 2009): we are empathetic by virtue of our intelligence and capacity to apprehend the mental life of a relatively wide range of creatures, but far more sympathetic to the closer group into which are born or choose to join, remaining neutral or hostile to “out-group” individuals.¹

Thus, in spite of our natural capacity for empathy, sympathy, and mutual assistance, the human being can also be described as a self-interested, control-oriented, dissociative xenophobe. In view of their historic prevalence, it is not unlikely that these features have evolved to become a part of our innate neurobiological identity and that any attempt to construe social structures (rules, conventions, contracts, etc.) opposing this identity must, in order to be realistically implemented, take this biological challenge into account in addition to the historically well-known political, social, and cultural challenges.

A major practical problem is that the effects of our actions are not limited, as are our capacities for engagement. The difficulty of wide involvement due to the brain's self-projective egocentricity is matched by a capacity to cause large-scale effects, which poses serious problems whenever large-scale or long-term solutions are needed—say, to improve the global environment, reduce global poverty, or safeguard future generations. Our societies are importantly construed around egocentric and short-term perspectives—political, economical, etc.—making it extremely difficult to put global or long-term thought and foresight into practice. This is of course only to be expected, since our brains'

¹ I am here discussing social attitudes in terms of subjective evaluators, but they can also be discussed in terms of non-conscious non-feeling units. Some current neuroscience literature may prefer to discuss the issue not from the point of view of subjective definitions but rather from the perspective of relevance detection and evaluation that is objectively observed.

neuronal architectures are engaged in social interactions and determine the social structures that we can and do develop.

However, our brain identity incorporates social influence. Culture and nature stand in a relationship of mutual causal influence: whilst the organisation of our brains in part determines who we are and what types of societies we develop, our social structures also have a strong impact on the brain's organisation; notably, they impact upon cultural imprints epigenetically stored in our brains. The genetic control over the brain's development is subject to epigenetic evolutionary processes; that is to say, to a coordinated and organised neuronal development that is the result of learning and experience and that is intermixed with the action of genes. The door to being epigenetically proactive is, accordingly, opened. In the following analysis of epigenesis by selective stabilisation of synapses I shall discuss the relationship between genotype and brain phenotype; the paradox of non-linear evolution between genome and brain complexity; the selection of cultural circuits in the brain during development; and the genesis and epigenetic transmission of cultural imprints.

3 Neuronal epigenesis

3.1 Genotype & brain phenotype: The paradox of non-linear evolution between genome & brain complexity

The comparison between what we presently know about human genomes and the brain phenotype raises the paradox of a non-linear evolution between the complexity of the genome and that of the brain ([Changeux 1985, 2012b](#)). From a molecular neurobiologist's perspective, the cognitive abilities and skills required for the highest functions of the human brain are built from a cascade of events driven by a "genetic envelope", which makes the difference between *Homo sapiens* and the human family's earliest ancestors, but which cannot be simply related to genome size, nor to the number of genes.

The total amount of DNA housed in the haploid genome is approximately 3.1 billion

base pairs, but no more than 20,000–25,000 gene sequences (1.2% of our genome code for exons—the DNA components of genes), and this number does not significantly differ from mouse to human. Moreover, the difference in full DNA sequences are very limited: between humans and chimpanzees they comprise no more than 4% of the genome. However, the total number of neurons in the human brain is in the order of 85 billion, compared to about 70 million in the brain of the mouse ([Azevedo et al. 2009](#)). Yet, notwithstanding the increase in cell numbers, with each neuron possessing its particular connectivity and its set of genes expressed, mammalian brain anatomy has evolved dramatically from a poorly corticalized lissencephalic brain with about 10–20 identified cortical areas to a brain with a very high relative cortical surface, multiple gyri and sulci, and possibly as many as 100 identified cortical areas ([Mountcastle 1998](#)). Thus, there exists a remarkable nonlinear relationship between the evolution of brain anatomy and the evolution of the genome organisation.

Molecular and cellular explanations have been suggested to account for this nonlinear relationship. One is the combinatorial expression of spatio-temporal patterns of genes that affect development ([Changeux 1985; Edelman 1987; Tsigelny et al. 2013](#)). Another, non-exclusive explanation, is the contribution of "epigenetic mechanisms" driven by interaction with the environment in the course of the long postnatal period of brain maturation—circa 15 years in humans—during which critical and reciprocal relationships take place between the brain and its physical, social, and cultural environment. It is on these epigenetic mechanisms that I shall focus here.

3.2 The epigenesis of neuronal networks by selective stabilization of synapses

The word "epigenesis" can be traced back to [William Harvey \(1651\)](#), who stated in contrast to contemporary preformationist views that the embryo arises by "the addition of parts budding out from one another". It was subsequently used by [Conrad Waddington \(1942\)](#) to specify the re-

lationship between the genes and their environment to produce a phenotype. This is also the meaning adopted in the theory of the epigenesis of neuronal networks by selective stabilization of synapses, according to which the environment affects the organisation of connections in an evolving neuronal network through the stabilization or elimination (pruning) of labile synapses, under the control of the state of activity of the network (Changeux et al. 1973). This meaning, which I shall use henceforth, contrasts with the more recent and biochemically distinct meaning of the word *epigenetic*, **which** refers to the status of DNA methylation and histone modification in a particular genomic region. This concerns the neuronal nucleus, but not the diversity of individual synaptic contacts (Sassone-Corsi & Christen 2012). The modulatory role of chromatin modifications in long-term memory has already been described (see e.g., Levenson & Sweatt 2005), but the informational content involved—which relies upon cell bodies—is expected to be in orders of magnitude smaller than that of synaptic epigenesis, based upon the combinatorial power of individual synapses.

During embryonic and postnatal development, the million billion (10^{15}) synapses that form the human brain network do not assemble like the parts of a computer, that is, according to a plan that precisely defines the disposition of all the individual components. If this were the case, the slightest error in the instructions for carrying out this program could have catastrophic consequences. On the contrary, the mechanism appears to rely on the progressive setting of robust interneuronal connections through trial-and-error mechanisms that formally resemble an evolutionary process by variation selection (Changeux et al. 1973; Changeux & Danchin 1976; Edelman 1987; Changeux 2012a). At sensitive periods of brain development, the phenotypic variability of nerve cell distribution and position, as well as the exuberant spreading and the multiple figures of transiently-formed connections originating from the erratic wandering of growth cone behaviour, introduce a maximal diversity of synaptic connections. This variability is then reduced by the selective stabilization of some of the labile con-

tacts and the elimination (or retraction) of others. The crucial hypothesis of the model is that the evolution of the connective state of each synaptic contact is governed globally, and within a given time window, by the overall “message” of signals experienced by the cell on which it terminates (Changeux et al. 1973).

One consequence of this is that particular electrical and chemical spatiotemporal patterns of activity in developing neuronal networks are liable to be inscribed under the form of defined and stable topologies of connections within the frame of the genetic envelope. In humans, about half of all adult connections are formed after birth at a very fast rate. The nesting of these multiple traces directly contributes to forming and shaping the micro- and macroscopic architecture of the wiring network of the adult human brain, thus bringing an additional explanation to the above-mentioned non-linearity paradox.

Another consequence of the synapse-selection model (originally presented as a “theorem of variability”) is that the selection of networks with different connective topologies can lead to the same input-output behavioural relationship (Changeux et al. 1973). This accounts for an important feature of the human brain: the constancy or “invariance” of defined states of behaviour, despite the epigenetic “variability” between individual brains’ connectivity.

Finally, both the spontaneous and the evoked activity may contribute to synapse selection. In this framework, a suggestion has been made that reward signals received from the environment may control the developmental evolution of connectivity (Gisiger et al. 2005; Gisiger & Kerszberg 2006). In other words, reinforcement learning would modulate the epigenesis of the network. The model has been implemented in a case of the learning of a visual delayed-matching-to-sample task (see below). This process of synaptic selection by reward signals may concern the evolution of brain connectivity in single individuals, but it also concerns the exchange of information and shared emotions or rewards between individuals in the social group (Changeux 1985, 2004; Gisiger et al. 2005). This is an important part of our argument; it may

thus play a critical role in social and cultural evolution.

3.3 The selection of cultural circuits in the brain during development & the epigenetic transmission of cultural imprints

There is an abundance of experimental studies that are consistent with, or directly support, the model of synapse selection. In humans the maximum synaptic density is reached within three years, then steadily declines until the total number stabilises around the time of puberty (Huttenlocher et al. 1997; Bourgeois 1997; Petanjek et al. 2011). Yet the process of synaptic refinement goes far beyond puberty: learning is life-long (Petanjek et al. 2011). The observed global decline in synaptic numbers during childhood plausibly reflects a rich cascade of elementary steps of learning by selection. Numerous studies have shown that when neuronal activity is experimentally modified, synaptic elimination is altered (Benoit & Changeux 1975, 1978; Stretavan et al. 1988; Purves & Lichtman 1980; Luo & O'Leary 2005; Innocenti & Price 2005; Collin & van den Heuvel 2013). At variance with the classical Lamarckist-constructivist scheme (Quartz & Sejnowski 1997), blocking the activity maintains a high number of connections: it is activity that enhances synaptic elimination (Benoit & Changeux 1975, 1978; Stretavan et al. 1988; Luo & O'Leary 2005). Thus “to learn is to eliminate” (Changeux 1985).

Among the cortical connections established in post-natal life are the long-range tracts between the frontal areas (Miller & Cohen 2001; Fuster 2008) and other brain cortical areas (including sensory ones) (Goldman-Rakic 1987; Goldman-Rakic 1999; Hagmann et al. 2008; Collin & van den Heuvel 2013). Some years ago, it was suggested, according to the “global neuronal workspace” hypothesis, that these long-range connections, by broadcasting signals to multiple brain areas, yield subjective “conscious” experience by allowing sensory inputs—seeing, hearing and so on—global access to many brain areas (Dehaene et al. 1998; Dehaene

& Changeux 2011). The long-range connections would provide a structural basis for the global experience known as conscious access.

These long-range connections are particularly important in the case of the prefrontal areas which contribute to planning, decision-making, thought, and socialisation. The ontogeny and postnatal development of long-range connectivity expectedly reveal phases of exuberance and phases of selection and axonal pruning (Collin & van den Heuvel 2013). In human newborns evolution is slow, and it has been suggested that the phase of exuberant long axon removal is largely completed at the age of two years, accompanied by increasing information processing and cognitive development (Collin & van den Heuvel 2013). Evolution continues during adolescence until adulthood with decreasing segregation and increasing integration, mainly but not exclusively driven by modulation of connections strength (local synaptic elimination persists in the adult; Petanjek et al. 2011). It is expected to have major consequences on the laying down of cultural imprints including the “epigenetic rules” associated with socialisation.

The acquisition of reading and writing may be viewed as a typical example of epigenetic development of “cultural circuits”. Writing and reading are recent cultural inventions (about 5000 years old) that evolved into distinct sub-systems and put considerable demands on our cognitive system. Historically, the first evidence for specialized writing and reading circuits in the brain was the discovery by the French neurologist Dejerine (1895) of pure alexia, also known as alexia without agraphia. Individuals with pure alexia suffer from severe reading problems while other language-related skills such as naming, oral repetition, auditory comprehension or writing are typically intact. Alexia results from cerebral lesions in circumscribed brain regions including the angular and supramarginal gyri. New specialized sets of connections are present exclusively in individuals that have learned written language and have been selected and consolidated in the course of development at sensitive periods (4–6 years) as a consequence of an intensive period of education.

The human brain did not evolve to learn to read, but possesses enough epigenetic variability in the course of its development (and also—though to a lesser extent—in the adult) to incorporate a cultural invention of this kind. During the acquisition of reading and writing by Western subjects, representations for visual forms of words progressively settle into the occipito-temporal cortex, recruiting a subset of functionally-appropriate object recognition regions in the temporo-parietal junction (Dehaene et al. 2010). The group of illiterate individuals is consistently more right-lateralized than their literate controls (Pettersson et al. 2007). Interestingly, alphabetic writing systems recruit circuits that differ in part from those mobilized by the Chinese ideographic systems. In French readers reading French, activations were enhanced in left-hemisphere visual area V1, with the strongest differences between French words and their controls found at the central and horizontal meridian representations. In contrast, Chinese readers reading Chinese showed enhanced activations in intermediate visual areas V3v/hV4, which was absent in French participants (Szwed et al. 2014). Also, the capacity to read sheet music is selectively altered in music-specific forms of alexia. Neuronal circuits specific to a given culture may thus become epigenetically established in the brains of social group members. Written language-learning is only one of the many cultural imprints acquired during the development of the human brain (Changeux 1985). For instance, cross-cultural differences between Asian and Western participants manifest themselves as differential increases of fMRI in the medial prefrontal cortex with reference to self-judgment (Zhu et al. 2007; Ray et al. 2010) and also to diverse brain recordings in mind reading (Kobayashi et al. 2007), holistic attention (Hedden et al. 2008), or facial photo recognition (Na & Kitayama 2011). The adult human brain thus builds up from a complex intertwining of cultural circuits progressively laid down during development within the framework of a human-specific genetic envelope.

There is no compelling evidence that culturally-acquired phenotypes will sooner or later

be genetically transmitted. What the evidence does show is that they have to be learned by each generation, by children from adults, and epigenetically transmitted from generation to generation, beginning in the mother's womb and up until the adulthood. Teaching reading and writing to circa five-year-old children requires elaborate pedagogic strategies, which in a general manner are absent in non-human primates (Premack 2007).

In short, cultural imprints have a physical reality in the human brain. Cultural imprints have also been demonstrated in non-human brains, e.g., by Peter Marler's work on birds' song-learning (Marler 1970). Yet the importance of cultural imprints on behaviour are comparatively much more important in humans compared to non-humans, in particular due to the long postnatal period of brain maturation. They play a critical role in shaping the brain phenotype in relation with the social group, through oral and written language but also through diverse culture-specific habits, traditions, and symbolic systems, including the ethical and social norms embodied in the adult brain.

I shall now proceed to discuss issues raised by the combinatorial explosion of brain representations and the channelling of behaviour through *epigenetic rules* and top-down control of decision-making.

epigenetic rules =_{Df} In neurobiological terms, these “rules” shall be viewed as acquired patterns of connections (scaffoldings), hypothetically stored in frontal cortex long-term memory. They frame the genesis of novel representations and regulate decision-making in a top-down manner.

4 “Epigenetic rules” and top-down control of decision-making

4.1 The hierarchical architecture of the brain

It has been suggested that ethical and social norms are, from a perspective in which the brain is central, ultimately encoded as spati-

otemporal patterns of neuronal activity that can be mobilized within the conscious neuronal workspace (Dehaene & Changeux 2011). Yet from a neurobiological standpoint, this view hinges upon the classical issue of the combinatorial explosion raised by the immense network of almost a million billion (10^{15}) interconnected synapses of the human brain. The question that arises, then, is how the particular patterns of neuronal activity, which, for instance, encode defined actions or perceptual events and ultimately ethical rules, are selected within this gigantic neural network. In my view, the concept of a hierarchical organisation of the brain needs to be taken into consideration more closely.

Analysis of the neurological deficits caused by lesions discloses hierarchical and parallel neural architectures that help us understand higher brain functions (Shallice & Cooper 2011). Among these is the inhibition of automatic (or reflex) actions and the elaboration of goal-directed behaviours and their control. In the brain, an evolutionary-recent territory of cerebral cortex architecture, the lateral prefrontal cortex, has been shown to play a critical role in the temporal control of behaviour. It serves as a “temporal buffer” between past events and future actions, allowing behaviours that follow internal goals to occur (Fuster 2001; Goldman-Rakic 1987; Petrides 2005). Moreover, the lateral prefrontal cortex exerts top-down control of cognitive processes associated with hierarchically-lower regions distributed in more posterior territories on the basis of internal plans, goals, or what may be referred to as “rules” (Miller & Cohen 2001; Passingham 1993; Shallice 1988; Dehaene & Changeux 1991; Koechlin et al. 2003). It thus contributes to decision-making within the actual context of a given individual history and stored memories (Damasio 1994) and to “neurally encoded rules” that can associate a context with a specific behavioural response and the ability to generalize a rule in novel circumstances.

An early formal model of learning by selection according to a rule was devised in the Wisconsin Card Sorting Task, which is commonly used as a test of the integrity of frontal lobe functions (Dehaene & Changeux 1991). It

requires subjects to infer a “rule” according to which a deck of cards must be sorted, i.e., colour, shape, or number. Feedback from the experimenter takes the form of a simple positive or negative reward (correct or incorrect). The goal for the subject is to get as many “right” responses as possible. Initially, cards must be sorted according to, say, colour. When performance is successful, the “sorting rule” is changed, for example from colour to shape; the subject must notice the change and find the new rule. The global architecture of a network that passes the task comprises two distinct levels of organization: a low level (level 1) that governs the orientation of the organism toward an object with a defined feature and which would correspond to a visuo-motor loop, including visual areas and the premotor cortex; and a high level (level 2) that controls the behavioural task according to a memory rule, and which would be homologous to the prefrontal cortex or closely-related areas (Dehaene et al. 1987; Dehaene & Changeux 1989).

A key feature of the model is that the high level contains a particular category or cluster of neurons, referred to as “rule-coding clusters”, each of which codes a single dimension (e.g., number, colour, or shape). During the acquisition step, the layer of rule-coding neurons is assumed to play the role of a “generator of diversity”. The spontaneous activity then plays a critical role in the activation of a given rule-coding cluster; and because of lateral inhibition only one cluster is active at a time. A search by trial and error takes place, until a positive reward is received from the environment (here the experimenter). Then, the particular cluster active at this precise moment is selected (for discussion see Monchi et al. 2001; Asplund et al. 2010; Fuster 2008). The number of trials necessary to learn the current rule is small (1–2), and single trial learning may occur in normal subjects as it does with the model (Dehaene & Changeux 1991). This learning of short-term rules based upon the fast (millisecond to second) allosteric transitions of synaptic receptors may also be transferred to long-term stores as epigenetically-acquired patterns of connections (see above).

In the course of the modelling of the Wisconsin card-sorting task, an additional architecture was introduced in the form of an auto-evaluation loop, which can short-circuit the reward input from the exterior. It allows for an internal evaluation of covert motor intentions without actualizing them as behaviours, but instead by testing them by comparison with memorized former experiences (Dehaene & Changeux 1991).

In these early formulations, the “rule-coding clusters” were pre-wired in the neuronal network. Subsequent models, however, opened the range of possible epigenetic rules to a brain-wide space of combinations made available within the global neuronal workspace (Baars 1988). This is of importance when we consider the ability to coordinate thoughts or actions in relation to internal goals, which is referred to as “cognitive control” and is a rather infrequent phenomenon. This discussion thus illustrates how rules encoding ethical norms may originate from the brain. Against this background—which shows how ethical rules might be epigenetically built from brain organization—I propose the possibility of being epigenetically proactive, and adapting our social structures, in both the short- and long-term, to benefit, influence, and constructively interact with the ever-developing neuronal architecture of our brains.

4.2 A cascade model of top-down cognitive control

Cognitive control has been further investigated by Koechlin et al. (2003) using a set of more complex tasks than the Wisconsin Card Sorting Task, and which span (at least three) nested levels of complexity. They consist in the presentation of series of coloured visual stimuli (squares or letters) organized into blocks, with an increasing importance of contextual signals: from “sensory control” with little if any contextual signal, to “contextual control” and, at the higher level, to “episodic control”. Brain imaging fMRI recordings with healthy human subjects revealed that the lateral prefrontal cortex contributes to a hierarchical cascade of executive processes that involve at least three nested

levels of processing. These are neurally implemented in distinct regions, from posterior premotor to rostral lateral prefrontal cortex regions (typically Brodman’s area 46; Koechlin et al. 2003; Badre & D’Esposito 2007; Badre et al. 2009). Patients with focal lateral prefrontal cortex lesions performed cognitive tasks with sensory, contextual, and episodic deficits associated with focal damage to Brodman’s areas 6, 45, and 46, respectively—as is expected from the cascade model (Azuar et al. 2014; Kayser & D’Esposito 2013).

By analogy with the Wisconsin Card Sorting Task (WCST) model mentioned above, behavioural rules are also sorted, but at different nested levels of information processing, the highest level rules “controlling” in a top-down manner the underlying rules closer to the senses. Hypothetically, ethical norms may be viewed as some particular kind of “control rules” developed within a social context, though this possibility still deserves to be explored by Koechlin, D’Esposito and colleagues.

Recently Collins & Koechlin (2012) have further suggested a computational model of human executive functioning associated with the prefrontal cortex, which integrates multiple processes during decision-making, such as expectedness of uncertainty, task switching, and reinforcement learning. The model reveals that the human frontal function may monitor up to three or four concurrent behavioural strategies and infers online their ability to predict action outcomes: whenever one appears more reliable than unreliable, this strategy is chosen to guide the selection and learning of actions that maximize rewards (see also Miller & Cohen 2001; Passingham 1993; Shallice 1988; Fuster 2008; Dehaene & Changeux 2011).

In their original paper, Collins and Koechlin do not explicitly mention social interaction. Yet we may consider an extension of their model to the social context by assuming that ethical or social norms are part of the “concurrent behavioural strategies” that they postulate exist in decision-making. The selection and learning of actions would then be more elaborate than the simple maximization of immediate rewards.

The developing baby is exposed very early on to a defined social and cultural environment, possibly even pre-natally (Lagercrantz & Changeux 2009; Lagercrantz et al. 2010). At this stage of development an intense synaptogenesis steadily occurs in the cerebral cortex, and epigenetic selection of neuronal networks accompanies the acquisition of the “maternal” language as well as of the common rules of the social community to which the child’s family belongs. The developing baby/child is “impregnated” with the current ethical rules of the social community, and this is often linked with the symbolic (philosophical/religious) system of representation character of the community to which it belongs. These early traces may last for the lifetime of the individual and sooner or later create conflicting relationships with a fast-evolving environment aggravated by the increased longevity of the individual (Changeux 1985). On the basis of the neurobiological data mentioned above, one may define these rules as epigenetically-acquired patterns of connections (scaffoldings) stored in frontal cortex long-term memory, which frame the genesis of novel representation and “cognitively controlled” decision-making in a top-down manner.

Against this background I propose the possibility of being epigenetically proactive and adapting our social structures, in both the short- and the long-term, to benefit, influence, and constructively interact with the ever-developing neuronal architecture of our brains.

5 A naturalistic responsibility

5.1 Proactive epigenesis

The first sentence in the 1948 Universal Declaration of Human Rights states: “All human beings are born free and equal in dignity and rights.”

Read as a description of the actual situation of human beings, this is blatantly and tragically false. Read as a normative ideal that we should strive for, it is noble but tragically unrealistic: considering our present cerebral structure, we are not likely to acknowledge in actual social practice the equal dignity and rights of all

individuals independently of race, gender, creed, etc. Life conditions may have improved for many humans over time, yet the present global situation remains appalling, notably, with respect to poverty, unequal distribution of health care, and the predominantly non-egalitarian or bellicose relations between individuals or groups. The vast majority of human beings appear reluctant, unable to identify with, or show compassion towards those who are beyond (and sometimes even towards those who are within) *their* sphere. While some societies or individuals may be more prone than others to developing a strong ethnic identity, violence, racism, sexism, social hierarchies, or exclusion, all exhibit some form and measure of xenophobia.

What I have here suggested, however, is that we might make presently unrealistic ideals, such as equality in dignity and rights, somewhat more realistic by selecting them for epigenetic proactivity.

Synaptic epigenetic theories of cultural and social imprinting on our brain architecture open the door to being epigenetically proactive, which means that we may culturally influence our brain organisation with the aim of self-improvement, individually as well as socially, and change our biological predispositions through a better fit of our brain to cultures and social structures.

I suggest that certain areas of research are especially important to pursue with the goal of “epigenetic proaction” in mind. They aim at integrating recent advances in neuroscientific research into normative debates at the level of society. This does not necessarily mean that my level of explanation is “neurocentric” or “neuroreductionist”. My aim is more “encyclopedic” in the sense that I wish to illustrate the benefits that neuroscience can bring to the humanities and social sciences and conversely. I do not see myself as either neuro-“centric” or “reductionist”—which would mean an exclusion of other categories of determinants at the social or historical levels—but I am more modestly willing to unify knowledge between the humanities and the neurosciences, which are too often deliberately omitted from the debate. This can be illustrated by two examples: violence in adoles-

cents in relation to their social environments, and violence in adults associated with interconfessional conflicts.

Violence in adolescents is a common phenomenon in our societies and it is frequently repressed through police and judiciary means, often resulting in incarceration. But this approach to juvenile violence simply omits the scientifically-established fact that adolescence is also a time of “neurodevelopmental crisis”. Evidence from anatomical and functional-imaging studies has highlighted major modifications of cortical circuits during adolescence. These include reductions of gyrification and grey matter, increases in the myelination of cortico-cortical connections, and changes in the architecture of large-scale cortical networks—including precentral, temporal, and frontal areas. (Klein et al. 2014). Uhlhaas et al. (2009) have used MEG synchrony as an indicator of conscious access and cognitive performance (rev. Dehaene & Changeux 2011). Until early adolescence, developmental improvements in cognitive performance are accompanied by increases in neural MEG synchrony. This developmental phase is followed by an unexpected decrease in neural synchrony that occurs during late adolescence and is associated with reduced performance. After this period of destabilization follows a reorganization of synchronization patterns that is accompanied by pronounced increases in gamma-band power and in theta and beta phase synchrony (Uhlhaas et al. 2009). These remarkable changes in neural connectivity and performance in the adolescent are only just being explored and may lead to special unexpected proactive care from society. In turn, this requires active research, including a social educative environment adequate to adolescents’ special needs. This may include adequate physical exercise, cultural games, educational training, and new kinds of therapies yet to be invented.

Violent interconfessional conflicts have raged throughout human history. They continue to plague our modern societies and are presently an important cause of wars and other forms of violence throughout the world. One should remember that every newborn and child brain incorporates critical features of its biolo-

gical, social, and cultural environment including, in addition to spoken and written language, symbolic systems and religious rituals (which include dietary and vestimentary practices as markers of the faith). These epigenetic traces are almost irreversibly laid down and may persist throughout the whole life of the individual. Yet they might be renewed through epigenetic transmission from adults to newborns. In this context, early proactive epigenetic imprinting through education is of critical importance. The aim of that education should not be to abolish faith or emotional convictions (e.g., moral, political, or religious) but only to control the fervour, intolerance, and fanaticism in their expression. The problem, as I see it, is not a belief itself, but the emotional intensity to which it gives rise and the manner in which it is expressed. Influencing a child brain to reduce its propensity to ideological violence or fanaticism and enhance its tolerance to others’ differences also requires special proactive care from society that per force involves active research—including a social educative environment adequate to this particular goal.

These are only two illustrations of the many that are possible, chosen because they have been problematic throughout the history of humankind and show no signs of disappearing.

At the individual level, the social conditions of an infant, or an adolescent, are of crucial importance in their cerebral development, and adequate conditions can in principle be provided. The factual realism of this application is largely a matter of political will and social agreement. The scientific challenge will be to further develop the knowledge of these conditions and their effects on the developing infant and adolescent brain. Also, the challenge will be to develop our knowledge of how social conditions affect the adult brain, e.g., to prevent neurodegeneration.

On a more general level, when applied on a larger scale to a society, a population, or to the entire human species, the argument follows the same logic and is no less important—but it becomes considerably more complicated to apply, theoretically as well as practically.

If new cultural imprints were epigenetically stored in our brains (say, less violent or less sectarian features), future generations would presumably develop societies that reflect them (i.e., become more peaceful and inclusive). A weakness of this optimistic reasoning is its circularity, since we would already need to be peaceful in order for a peaceful society to be maintained. A crucial question then becomes: how long does it take for a cultural characteristic to leave a cerebral trace? In some measure stable and enduring cultural structures are needed in order to effect stable neurobiological changes and store cultural imprints in the brain that might give evolution a push in the right direction, but the chances of maintaining societies that conflict with the present nature of its inhabitants—say, maintaining a peaceful egalitarian rule in a society of violent xenophobes—are arguably slim.

The challenges involved in trying to be epigenetically proactive by culturally influencing the future actions of human genes and neuronal structures, with the aim of altering higher cognitive functions and their resulting behaviour seem formidable, at least if enlarged sympathy is on the agenda. Still, within the epigenetic neuroscientific framework, at least the theoretical possibility exists, and it is worthy of consideration by many other disciplines beyond neuroscience. Depending on how we choose to develop our culture, one day epigenetic rules that enlarge the presently-narrow realm of human sympathy might perhaps emerge.

5.2 Conclusion: A naturalistic responsibility

The origins of norms and the relationship between facts and values have been much debated in philosophy. Reasoning that weds scientific theory with normative considerations has been accused of committing the logical error of confusing facts and values, which is known as “the naturalistic fallacy”.

The expression “the naturalistic fallacy” was coined by the British moral philosopher G. E. Moore and refers in his work to the identification (or reduction) of goodness with (or to)

another property such as utility, pleasure, or happiness (Moore 1903). That issue is not relevant in the present context. In the interpretation of the naturalistic fallacy that is relevant here, the fallacy consists in deriving an “ought” from an “is”, or a value from a fact, and letting descriptive properties entail normative properties, which confuses the distinction between facts and values in a fallacious manner. This argument is reminiscent of David Hume’s claim that what *is* is entirely different from what *ought to be*, for “the distinction of vice and virtue is not founded on the relations of objects, nor is perceiv’d by reason” but is fundamentally a matter of feelings and as such is neither true nor false (Hume 1739, III, I). I agree that it is fallacious to derive “ought to be” from “is”, and consider this a conceptual mistake that our theory of epigenetic proaction must and indeed does avoid. I do not assert that factual descriptions of the brain’s architecture are tantamount to yielding recommendations or assertions of norms, do not confuse “is” with “ought”, and consequently do not commit the naturalistic fallacy in this formulation.

We should observe that a *value* may be represented on many levels: non-conscious as well as conscious, as a basic biological function or as a feature of advanced moral reasoning. When discussing the naturalistic fallacy, value as a feature of advanced normative reasoning is the relevant sense of the term. The logical distinction between fact and value could collapse if the term is defined differently—say, if it features as a non-normative biological function. The logical error in the naturalistic fallacy concerns the fact/value distinction as it is drawn between normative and descriptive statements, namely between *ought* and *is*; not between facts that are/are not biological values, where that concern would presumably not arise.

However, eagerness to avoid the naturalistic fallacy must not prevent our normative reasoning from being informed by scientific theories. Normative judgments should be informed by facts, even though they cannot be entailed by them. If certain evaluative tendencies are innate in the normal human brain’s architecture, such as self-interest and selective sympathy, this

fact (if it is one) about the human being's neuronal structure would admittedly entail that every healthy, sufficiently mature individual will to some degree feel both self-interest and sympathy towards some other creature. However, this is not the entailment of a norm, but an empirical entailment of another fact. It does not entail that it is good (or bad), or that we ought to conceive it as good (or bad) that we are thus construed. Similarly, if it is true that we are, for example, and as we have argued, self-projective xenophobes, knowledge of this (presumed) fact is not in itself a justification of it. Understanding is not the same as justification: to know, or to understand, is not to approve. On the contrary, knowledge about our neural structures' predispositions should increase our awareness of the need for stable and realistic social structures and agreements to keep us in check.

We should also observe that a belief in the approximate universality of certain values, or preferential tendencies as innate features of the human neurobiological make-up, is logically compatible with a belief in maintaining the description/norm distinction.

My primary focus has been on the important empirical connections between biological facts and norms. Norms are brain constructs elaborated by human societies, biologically as well as culturally embedded in and constrained by the contingent evolution of socio-cultural structures—in particular, by the multiple symbolic philosophical and religious systems that have developed. This fact, and the realisation that normative judgments should be informed by facts even though they cannot be entailed by them suggests that science, philosophy and—not least—neuroethics—have a major responsibility: namely to decipher the network of causal connections between the neurobiological, socio-cultural, and contingent historical perspectives that allow a moral norm to be enunciated at a given moment in human history; and to evaluate their “universal” character as pre-specified in our genome and shared by the human species in distinction from those relative to a given culture or symbolic system. The “fallacy” of the naturalistic approach is thus inverted into a naturalistic responsibility (Evers 2009): the re-

sponsibility to connect facts and values, biology, and socio-cultural structures, and to use that enriched understanding for the benefit of ourselves and our societies.

We may hope that through the rational exchange of arguments between partners with different cultures and moral traditions debating together, a species-specific “human core” could become dominant beyond individual differences and converge on a common structure (Changeux & Ricoeur 2000). At the same time, we must note that the diversity of human individuals and societies is enormous and must be respected while we strive to find this common ground that might allow coexistence.

The idea of proactively selecting those specific dispositions or capacities (such as sympathy) that we all share as human beings which that, if properly developed, may benefit our global co-existence while respecting individual and ideological diversities, is well in line with Darwin. Darwin wrote in *The Descent of Man*:

As man advances in civilization, and small tribes are united into larger communities, the simplest reason would tell each individual that he ought to extend his social instincts and sympathies to all members of the same nation, though personally unknown to him. This point being once reached, there is only an artificial barrier to prevent his sympathies extending to the men of all nations and races.

Lewontin (1993) argues that while traditional Darwinism has portrayed the organism as a passive recipient of environmental influences, a correct understanding should emphasize that humans are active constructors of their own environment—in particular the social and cultural environment. I agree and argue further that, in line with Darwin, we can be active constructors of our own brains through using our environment and culture, in a relationship that is reciprocal.

In this article, my main focus has been on feasibility—that is, on whether we *can* be epigenetically proactive. If we assume an affirmative answer to that question, an important fol-

low-up question arises: whether we *should* be so. My basic position, that I have here tried to express, is that epigenetic proaction could be a very promising, powerful, and long-term way of influencing human nature and of improving our societies. However, in order to pursue this in a responsible and adequate manner, caution is required, along with careful analyses of the relevant social and ethical issues. Science can be, and has throughout history repeatedly been, ideologically hijacked, and the resulting dangers increase with the strength of the science in question. If, say, humans learn to design their own brain more potently than we already do by selecting what we believe to be brain-nourishing food and pursuing neuronally-healthy life-styles, we *could* use that knowledge well—that is, there is certainly room for improvement. On the other hand, the dream of the perfect human being has a sordid past, providing ample cause for concern about such projects. Historic awareness is of the utmost importance for neuroethics when assessing suggested applications in a responsible and adequate manner. Moreover, what we mean by “responsible and adequate” is open to interpretation. The traits we choose to favour epigenetically, and the social structures we choose to develop, depend on who “we” are, and in what society we wish to live.

Arthur Koestler compares evolution to “a labyrinth of blind alleys” and suggests that “there is nothing very strange or improbable in the assumption that man’s native equipment, though superior to that of any other living species, nevertheless contains some built-in error or deficiency which predisposes him to self-destruction” (Koestler 1967, xi). In that light, steering evolution by influencing the cultural imprints to be stored in our brains appears to be an attractive option.

Acknowledgements

I wish to thank Jean-Pierre Changeux for his important scientific contributions to this paper, and for his detailed scrutiny of the arguments expressed. I also wish to thank Yadin Dudai, Sten Grillner, Hugo Lagercrantz, and Arleen Salles for their valuable comments on earlier

versions of this manuscript. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 604102 (HBP).

References

- Asplund, C. L., Todd, J. J., Snyder, A. P., Gilbert, C. M. & Marois, R. (2010). Surprise-induced blindness: A stimulus-driven attentional limit to conscious perception. *Journal of Experimental Psychology Human Perception and Performance*, 36 (6), 1372-1381. [10.1037/a0020551](https://doi.org/10.1037/a0020551)
- Azevedo, F. A. C., Carvalho, L. R. B., Grinberg, L. T., Farfel, J. M., Ferretti, R. E. L., Leite, R. E. P., Filho, J. W., Lent, R. & Herculano-Houzel, S. (2009). Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *Journal of Comparative Neurology*, 513 (5), 532-541. [10.1002/cne.21974](https://doi.org/10.1002/cne.21974)
- Azuar, C., Reyes, P., Slachevsky, A., Volle, E., Kinkingnehun, S., Kouneiher, F., Bravo, E., Dubois, B., Koechlin, E. & Levy, R. (2014). Testing the model of caudorostral organization of cognitive control in the human with frontal lesions. *NeuroImage*, 1 (84), 1053-60-1060. [10.1016/j.neuroimage.2013.09.031](https://doi.org/10.1016/j.neuroimage.2013.09.031)
- Baars, B. J. (1988). *A cognitive theory of consciousness*. Cambridge, UK: Cambridge University Press.
- Badre, D. & D'Esposito, M. (2007). Functional magnetic resonance imaging evidence for a hierarchical organization of the prefrontal cortex. *Journal of Cognitive Neuroscience*, 19 (12), 2082-2099. [10.1162/jocn.2007.19.12.2082](https://doi.org/10.1162/jocn.2007.19.12.2082)
- Badre, D., Hoffman, J., Cooney, J. W. & D'Esposito, M. (2009). Hierarchical cognitive control deficits following damage to the human frontal lobe. *Nature Neuroscience*, 12 (4), 515-522. [10.1038/nn.2277](https://doi.org/10.1038/nn.2277)
- Barto, A. G. & Sutton, R. S. (1982). Simulation of anticipatory responses in classical conditioning by a neuron-like adaptive element. *Behavioral Brain Research*, 4 (3), 221-235. [10.1016/0166-4328\(82\)90001-8](https://doi.org/10.1016/0166-4328(82)90001-8)
- Benoit, P. & Changeux, J.-P. (1975). Consequences of tenotomy on the evolution of multiinnervation on developing rat soleus muscle. *Brain Research*, 99 (2), 354-358.
- (1978). Consequences of blocking the nerve with a local anaesthetic on the evolution of multiinnervation at the regenerating neuromuscular-junction of the rat. *Brain Research*, 149 (1), 89-86.
- Berger, H. (1929). "Über das Enkephalogramm beim Menschen" [On the use of the encephalogram in humans]. *Archiv für Psychiatrie und Nerven-krankheiten*
- Blakemore, S. J., Wolpert, D. M. & Frith, C. D. (1998). Central cancellation of self-produced tickle sensation. *Nature Neuroscience*, 1 (7), 10196573-10196573. [10.1038/2870](https://doi.org/10.1038/2870)
- Bourgeois, J.-P. (1997). Synaptogenesis, heterochrony and epigenesis in the mammalian neocortex. *Acta Paediatrica Supplement* 42227-33
- Changeux, J.-P. (1985). *Neuronal man*. New York, NY: Pantheon Books.
- (2004). *The physiology of truth: Neuroscience & human knowledge*. Boston, MA: Harvard University Press.
- (2012a). Synaptic epigenesis and the evolution of higher brain functions. In P. Sassone-Corsi & Y. Christen (Eds.) *Epigenetics, brain and behavior* (pp. 11-22). Dordrecht, NL: Springer.
- (2012b). *The good, the true, the beautiful: A neuronal approach*. New Haven, CT: Yale/Odile Jacob.
- Changeux, J.-P. & Danchin, A. (1976). Selective stabilization of developing synapses as a mechanism for the specification of neuronal networks. *Nature*, 264 (5588), 705-712. [10.1038/264705a0](https://doi.org/10.1038/264705a0)
- Changeux, J.-P., Courge, P. & Danchin, A. (1973). A theory of the epigenesis of neuronal networks by selective stabilization of synapses. *Proceedings of the National Academy of Sciences USA*, 70 (10), 2974-2978.
- Changeux, J.-P. & Lou, H. C. (2011). Emergent pharmacology of conscious experience: New perspectives in substance addiction. *Federation of American Societies of Experimental Biology Journal*, 25 (7), 2098-2108. [10.1096/fj.11-0702ufm](https://doi.org/10.1096/fj.11-0702ufm)
- Changeux, J.-P. & Ricoeur, P. (2000). *What makes us think? A neuroscientist and a philosopher argue about ethics, human nature and the brain*. Princeton, NJ: Princeton University Press.
- Cheng, Y., Lee, P., Yang, C. Y., Lin, C. P. & Decety, J. (2008). Gender differences in the mu rhythm of the human mirror-neuron system. *PLoS ONE*, 3 (5), e2113-e2113. [10.1371/journal.pone.0002113](https://doi.org/10.1371/journal.pone.0002113)
- Collins, A. & Koechlin, E. (2012). Reasoning, learning, and creativity: Frontal lobe function and human decision-making. *PLoS Biology*, 10 (3), e1001293-e1001293. [10.1371/journal.pbio.1001293](https://doi.org/10.1371/journal.pbio.1001293)
- Collin, G. & van den Heuvel, M. P. (2013). The ontogeny of the human connectome: Development and dynamic changes of brain connectivity across the life span. *Neuroscientist*, 19 (6), 616-628. [10.1177/1073858413503712](https://doi.org/10.1177/1073858413503712)
- Damasio, A. (1994). *Descartes error: Emotion, reason, and the human brain*. New York, NY: Putnam.
- (1999). *The feeling of what happens: Body and*

- emotion in the making of consciousness*. San Diego, CA: Harcourt.
- Damasio, A. & Carvalho, G. B. (2013). The nature of feelings: Evolutionary and neurobiological origins. *Nature Reviews Neuroscience*, 14 (2), 143-152. [10.1038/nrn3403](#)
- Darwin, C. (1871). *The descent of man*. London, UK: John Murray.
- Decety, J. (Ed.) (2012). *Empathy: From bench to bedside*. Cambridge, MA: MIT Press.
- Decety, J. & Sommerville, J. A. (2003). Shared representations between self and others: A social cognitive neuroscience view. *Trends in Cognitive Sciences*, 7 (12), 527-533. [10.1016/j.tics.2003.10.004](#)
- Dehaene, S. & Changeux, J.-P. (1989). A simple model of prefrontal cortex function in delayed-response tasks. *Journal of Cognitive Neuroscience*, 1 (3), 244-261. [10.1162/jocn.1989.1.3.244](#)
- (1991). The Wisconsin card sorting test: Theoretical analysis and simulation of a reasoning task in a model neuronal network. *Cerebral Cortex*, 1 (1), 62-79.
- (2000). Reward-dependent learning in neuronal networks for planning and decision making. *Progress in Brain Research*, 126. [10.1016/S0079-6123\(00\)26016-0](#)
- (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, 70 (2), 200-227. [10.1016/j.neuron.2011.03.018](#)
- Dehaene, S., Changeux, J.-P. & Nadal, J.-P. (1987). Neural networks that learn temporal sequences by selection. *Proceedings of the National Academy of Sciences, USA*, 84 (9), 2727-2731.
- Dehaene, S., Kerszberg, M. & Changeux, J.-P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences USA*, 95 (24), 14529-14534. [10.1073/pnas.95.24.14529](#)
- Dehaene, S., Pegado, F., Braga, L. W., Ventura, P., Nunes Filho, G., Jobert, A., Dehaene-Lambertz, G., Kolinsky, R., Morais, J. & Cohen, L. (2010). How learning to read changes the cortical networks for vision and language. *Science*, 330 (6009), 1359-1364. [10.1126/science.1194140](#)
- Dejerine, J. (1895). *Anatomie des centres nerveux vol. 1*. Paris, FR: Rueffet Cie.
- Denton, D. (2006). *The primordial emotions: The dawn-ing of consciousness*. Paris, FR: Flammarion.
- Dudai, Y. (1989). *The neurobiology of memory: Concepts, findings, trends*. Oxford, UK: Oxford University Press.
- (2002). *Memory from A to Z: Keywords, concepts and beyond*. Oxford, UK: Oxford University Press.
- Edelman, G. (1987). *Neural Darwinism: The theory of neuronal group selection*. New York, NY: Basic Books.
- (1992). *Bright air, brilliant fire: On the matter of the mind*. New York, NY: Basic Books.
- Engen, H. G. & Singer, T. (2013). Empathy circuits. *Current Opinion in Neurobiology*, 23 (2), 275-282. [10.1016/j.conb.2012.11.003](#)
- Evers, K. (2001). The Importance of being a self. *International Journal of Applied Philosophy*, 15 (1), 65-83. [10.5840/ijap20011512](#)
- (2009). *Neuroéthique. Quand la matière s'éveille*. Paris, FR: Éditions Odile Jacob.
- Falck-Ytter, T., Gredebäck, G. & von Hofsten, C. (2006). Infants predict other people's action goals. *Nature Neuroscience.*, 9 (7), 878-879. [10.1038/nn1729](#)
- Fuster, J. M. (2001). The prefrontal cortex—An update: Time is of the essence. *Neuron*, 30 (2), 319-333. [10.1016/S0896-6273\(01\)00285-9](#)
- (2008). *The prefrontal cortex*. London, UK: Academic Press.
- Gallagher, S. (2000). Philosophical conceptions of the self: Implications for cognitive science. *Trends in Cognitive Sciences*, 4 (1), 14-21. [10.1016/S1364-6613\(99\)01417-5](#)
- Galli, L. & Maffei, L. (1988). Spontaneous impulse activity of rat retinal ganglion cells in prenatal life. *Science*, 242 (4875), 90-91. [10.1126/science.3175637](#)
- Gardiner, J. M. (2001). Episodic memory and autonoetic consciousness: A first-person approach. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 356 (1413), 1351-1361. [10.1098/rstb.2001.0955](#)
- Gisiger, T., Kerszberg, M. & Changeux, J.-P. (2005). Acquisition and performance of delayed-response tasks: A neural network model. , 15 (5), 489-506. [10.1093/cercor/bhh149](#)
- Gisiger, T. & Kerszberg, M. (2006). A model for integrating elementary neural functions into delayed-response behavior. *PLoS Computational Biology*, 2 (4), e25-e25. [10.1371/journal.pcbi.0020025](#)
- Goldman-Rakic, P. (1987). Circuitry of primate prefrontal cortex and regulation of behavior by representational memory. In F. Plum (Ed.) *Handbook of physiology* (pp. 373-417). Washington DC: The American Physiological Society.
- (1999). The physiological approach: Functional architecture of working memory and disordered cognition in schizophrenia. *Biological Psychiatry*, 46 (5), 650-661. [10.1016/S0006-3223\(99\)00130-4](#)
- Goodman, C. S. & Shatz, C. J. (1993). Developmental

- mechanisms that generate precise patterns of neuronal connectivity. *Cell*, 72, 77-98.
[10.1016/S0092-8674\(05\)80030-3](https://doi.org/10.1016/S0092-8674(05)80030-3)
- Gordon, G. & Ahissar, E. (2012). Hierarchical curiosity loops and active sensing. *Neural Networks*, 32, 119-129.
[10.1016/j.neunet.2012.02.024](https://doi.org/10.1016/j.neunet.2012.02.024)
- Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C. J., Wedeen, V. J. & Sporns, O. (2008). Mapping the structural core of human cerebral cortex. *PLoS Biology*, 16 (7), e159-e159.
[10.1371/journal.pbio.0060159](https://doi.org/10.1371/journal.pbio.0060159)
- Hart, A. J., Whalen, P. J., Shin, L. M., McInerney, S. C., Fischer, H. & Rauch, S. L. (2000). Differential response in the human amygdala to racial outgroup vs ingroup face stimuli. *NeuroReport*, 11 (11), 2351-2355.
- Harvey, W. (1651). *Exercitationes de generatione*. London, UK: Animalium.
- Hedden, T., Ketay, S., Aron, A., Markus, H. R. & Gabrieli, J. D. (2008). Cultural influences on neural substrates of attentional control. *Psychological Science*, 19 (1), 12-17. [10.1111/j.1467-9280.2008.02038.x](https://doi.org/10.1111/j.1467-9280.2008.02038.x)
- Hills, P. J. & Lewis, M. B. (2006). Reducing the own-race bias in face recognition by shifting attention. *Quarterly journal of experimental psychology*, 59 (6), 996-1002.
[10.1080/17470210600654750](https://doi.org/10.1080/17470210600654750)
- Hommel, B., Müsseler, J., Aschersleben, G. & Prinz, W. (2001). The Theory of Event Coding (TEC): A framework for perception and action planning. *Behavioral and Brain Science*, 849-937.
- Hume, D. (1739). *Treatise of Human Nature*.
- Huttenlocher, P. & Dabholkar, A. (1997). Regional difference in synaptogenesis in human cerebral cortex. *Journal of Comparative Neurology*, 387 (2), 167-178.
[10.1002/\(SICI\)1096-9861\(19971020\)387](https://doi.org/10.1002/(SICI)1096-9861(19971020)387)
- Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C. & Rizzolatti, G. (2005). Grasping the intention with one's own mirror neuron system. *PLoS Biology*, 3 (3), e79-e79.
[10.1371/journal.pbio.0030079](https://doi.org/10.1371/journal.pbio.0030079)
- Innocenti, G. M. & Price, D. J. (2005). Exuberance in the development of cortical networks. *Nature Reviews Neuroscience*, 6 (12), 955-965. [10.1038/nrn1790](https://doi.org/10.1038/nrn1790)
- Jackson, P. L. & Decety, J. (2004). Motor cognition: A new paradigm to study self-other interactions. *Current Opinion in Neurobiology*, 14 (2), 1-5.
[10.1016/j.conb.2004.01.020](https://doi.org/10.1016/j.conb.2004.01.020)
- Jackson, P. L., Brunet, E., Meltzoff, A. N. & Decety, J. (2006). Empathy examined through the neural mechanisms involved in imagining how I feel versus how you feel pain: An event-related fMRI study. *Neuropsychologia*, 44 (5), 752-761.
[10.1016/j.neuropsychologia.2005.07.015](https://doi.org/10.1016/j.neuropsychologia.2005.07.015)
- Jeannerod, M. (2006). *Motor cognition: What actions tell the self*. Oxford, UK: Oxford University Press.
- Kayser, A. S. & D'Esposito, M. (2013). Abstract rule learning: The differential effects of lesions in frontal cortex. *Cerebral Cortex*, 23 (1), 230-240.
[10.1093/cercor/bhs013](https://doi.org/10.1093/cercor/bhs013)
- Kitayama, S. & Uskul, A. K. (2011). Culture, mind, and the brain: Current evidence and future directions. *Annual Reviews of Psychology*, 62, 419-449.
[10.1146/annurev-psych-120709-145357](https://doi.org/10.1146/annurev-psych-120709-145357)
- Klein, D., Rotarska-Jagiela, A., Genc, E., Sritharan, S., Mohr, H., Roux, F., Han, C. E., Kaiser, M., Singer, W. & Uhlhaas, P. J. (2014). Adolescent Brain maturation and cortical folding: Evidence for reductions in gyrification. *PLoS ONE*, 9 (1), e84914-e84914.
[10.1371/journal.pone.0084914](https://doi.org/10.1371/journal.pone.0084914)
- Kobayashi, C., Glover, G. H. & Temple, E. (2007). Cultural and linguistic effects on neural bases of "theory of mind" in American and Japanese children. *Brain Research*, 1164, 95-107. [10.1016/j.brainres.2007.06.022](https://doi.org/10.1016/j.brainres.2007.06.022)
- Koechlin, E., Ody, C. & Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. *Science*, 302 (5648), 1181-1185.
[10.1126/science.1088545](https://doi.org/10.1126/science.1088545)
- Koestler, A. (1967). *The Ghost in the Machine*. UK: Arkana Books.
- Lagercrantz, H. (2005). *I barnets hjärna, Le cerveau de l'enfant*. Paris, FR: Éditions Odile Jacob.
- Lagercrantz, H. & Changeux, J.-P. (2009). The emergence of human consciousness: From fetal to neonatal life. *Pediatric Research*, 65 (3), 255-260.
[10.1203/PDR.0b013e3181973b0d](https://doi.org/10.1203/PDR.0b013e3181973b0d)
- Lagercrantz, H., Hanson, M., Ment, L. & Peebles, D. (Eds.) (2010). *The newborn brain neuroscience and clinical applications*. Cambridge, UK: Cambridge University Press, 2nd Edition.
- Lawrence, E. J., Shaw, P., Giampietro, V. P., Surguladze, S., Brammer, M. J. & David, A. S. (2006). The role of 'shared representations' in social perception and empathy: An fMRI study. *NeuroImage*, 29 (4), 1173-1184.
[10.1016/j.neuroimage.2005.09.001](https://doi.org/10.1016/j.neuroimage.2005.09.001)
- Ledoux, J. (1998). *The emotional brain: The mysterious underpinnings of emotional life*. Cambridge, UK: Cambridge University Press.
- Levenson, J. M. & Sweatt, J. D. (2005). Epigenetic mechanisms in memory formation. *Nature Reviews Neuros-*

- science, 6 (2), 108-118. [10.1038/nrn1604](https://doi.org/10.1038/nrn1604)
- Lewontin, R. (1993). *The doctrine of DNA: Biology as ideology*. London, UK: Penguin Books.
- Liberman, A. M. & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, 21 (1)
- Llinas, R. R. & Paré, D. (1991). Of dreaming and wakefulness. *Neuroscience*, 44 (3), 521-535. [10.1016/0306-4522\(91\)90075-Y](https://doi.org/10.1016/0306-4522(91)90075-Y)
- Lorenz, K. (1963). *On aggression*. London, UK: Methuen.
- Lou, H. C., Luber, B., Crupain, M., Keenan, J. P., Nowak, M., Kjaer, T. W., Sackeim, H. A. & Lisanby, S. H. (2004). Parietal cortex and representation of the mental self. *Proceedings of the National Academy of Sciences USA*, 101 (17), 6827-6832. [10.1073/pnas.0400049101](https://doi.org/10.1073/pnas.0400049101)
- Luo, L. & O'Leary, D. M. (2005). Axon retraction and degeneration in development and disease. *Annual Reviews of Neuroscience*, 28, 127-156. [10.1146/annurev.neuro.28.061604.135632](https://doi.org/10.1146/annurev.neuro.28.061604.135632)
- Marler, P. (1970). A comparative approach to vocal learning: Song development in white-crowned sparrows. *Journal of Comparative & Physiological Psychology*, 71 (2), 1-25. [10.1037/h0029144](https://doi.org/10.1037/h0029144)
- Michel, C., Rossion, B., Han, J., Chung, C. S. & Caldara, R. (2006). Holistic processing is finely tuned for faces of one's own race. *Psychological Science*, 17 (7), 608-615. [10.1111/j.1467-9280.2006.01752.x](https://doi.org/10.1111/j.1467-9280.2006.01752.x)
- Miller, E. K. & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Reviews of Neuroscience*, 24, 167-202. [10.1146/annurev.neuro.24.1.167](https://doi.org/10.1146/annurev.neuro.24.1.167)
- Monchi, O., Petrides, M., Petre, V., Worsley, K. & Dagher, A. (2001). Wisconsin card sorting revisited: Distinct neural circuits participating in different stages of the task identified by event-related functional magnetic resonance imaging. *Journal of Neuroscience*, 21 (19), 7733-7741.
- Moore, G. E. (1903). *Principia ethica*. Cambridge, UK: Cambridge University Press.
- Mountcastle, V. (1998). *Perceptual neuroscience: The cerebral cortex*. Cambridge, MA: Harvard University Press.
- Na, J. & Kitayama, S. (2011). Spontaneous trait inference is culture-specific : Behavioral and neural evidence. *Psychological Science*, 22 (8), 1025-1032. [10.1177/0956797611414727](https://doi.org/10.1177/0956797611414727)
- Narayanan, C. H. & Hamburger, V. (1971). Motility in chick embryos with substitution of lumbosacral by brachial by lumbosacral spinal cord segments. *Journal of Experimental Zoology*, 178 (4), 415-413. [10.1002/jez.1401780402](https://doi.org/10.1002/jez.1401780402)
- Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotions*. New York, NY: Oxford University Press.
- Parr, L. A. & Waller, B. M. (2006). Understanding chimpanzee facial expression: Insights into the evolution of communication. *Social Cognitive and Affective Neuroscience*, 1 (3), 221-228. [10.1093/scan/nsl031](https://doi.org/10.1093/scan/nsl031)
- Passingham, R. (1993). *The frontal lobes and voluntary action*. Oxford, UK: Oxford University Press.
- Petanjek, Z., Judas, M., Simic, G., Rasin, M. R., Uylings, H. B., Rakic, P. & Kostovic, I. (2011). Extraordinary neoteny of synaptic spines in the human prefrontal cortex. *Proceedings of the National Academy of Sciences, USA*, 108 (32), 13281-13286. [10.1073/pnas.1105108108](https://doi.org/10.1073/pnas.1105108108)
- Petersson, K. M., Silva, C., Castro-Caldas, A., Ingvar, M. & Reis, A. (2007). Literacy: A cultural influence on functional left-right differences in the inferior parietal cortex. *European Journal of Neuroscience*, 26 (3), 791-799. [10.1111/j.1460-9568.2007.05701.x](https://doi.org/10.1111/j.1460-9568.2007.05701.x)
- Petrides, M. (2005). Lateral prefrontal cortex: Architectonic and functional organization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360 (1456), 781-795. [10.1098/rstb.2005.1631](https://doi.org/10.1098/rstb.2005.1631)
- Phelps, E. A., Cannistraci, C. J. & Cunningham, W. A. (2003). Intact performance on an indirect measure of race bias following amygdala damage. *Neuropsychologia*, 41 (2), 203-208. [10.1016/S0028-3932\(02\)00150-1](https://doi.org/10.1016/S0028-3932(02)00150-1)
- Premack, D. (2007). Human and animal cognition: Continuity and discontinuity. *Proceedings of the National Academy of Sciences USA*, 104 (35), 13861-13867. [10.1073/pnas.0706147104](https://doi.org/10.1073/pnas.0706147104)
- Purves, D. & Lichtman, J. (1980). Elimination of synapses in the developing nervous system. *Science*, 210 (4466), 153-157. [10.1126/science.7414326](https://doi.org/10.1126/science.7414326)
- Putnam, F. (1989). *Diagnosis & treatment of multiple personality disorder*. London, UK: The Guilford Press.
- Quartz, S. R. & Sejnowski, T. J. (1997). The neural basis of cognitive development: A constructivist manifesto. *Behavioral Brain Sciences*, 20 (4), 537-596.
- Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A. & Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences USA*, 98 (2), 676-682. [10.1073/pnas.98.2.676](https://doi.org/10.1073/pnas.98.2.676)
- Ray, R. D., Shelton, A. L., Hollon, N. G., Matsumoto, D., Frankel, C. B., Gross, J. J. & Gabrieli, J. D. E. (2010).

- Interdependent self-construal and neural representations of the self and mother. *Social Cognitive and Affective Neuroscience*, 5 (2-3), 318-323. [10.1093/scan/nsp039](https://doi.org/10.1093/scan/nsp039)
- Ricoeur, P. (1992). *Oneself as Another* (K. Blamey trans). Chicago, IL: University of Chicago Press.
- Rochat, P. (2001). *The infant's world*. Cambridge, MA: Harvard University Press.
- Sassone-Corsi, P. & Christen, Y. (Eds.) (2012). *Epigenetics, Brain and Behavior*. Springer.
- Schultz, W. (2006). Behavioral theories and the neurophysiology of reward. *Annual Reviews in Psychology*, 57, 87-115. [10.1146/annurev.psych.56.091103.070229](https://doi.org/10.1146/annurev.psych.56.091103.070229)
- Schultz, W., Dayan, P. & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275 (5306), 1593-1599. [10.1126/science.275.5306.1593](https://doi.org/10.1126/science.275.5306.1593)
- Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge, UK: Cambridge University Press.
- Shallice, T. & Cooper, R. (2011). *The organisation of mind*. Oxford, UK: Oxford University Press.
- Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R. J. & Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science*, 303 (5661), 1157-1162. [10.1126/science.1093535](https://doi.org/10.1126/science.1093535)
- Singer, T., Seymour, B., O'Doherty, J. P., Stephan, K. E., Dolan, R. J. & Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature*, 439 (7075), 466-469. [10.1038/nature04271](https://doi.org/10.1038/nature04271)
- Stretavan, D. W., Shatz, C. J. & Stryker, M. P. (1988). Modification of retinal ganglion cell axon morphology by prenatal infusion of tetrodotoxin. *Nature*, 336 (6198), 468-471. [10.1038/336468a0](https://doi.org/10.1038/336468a0)
- Szwed, M., Qiao, E., Jobert, A., Dehaene, S. & Cohen, L. (2014). Effects of literacy in early visual and occipitotemporal areas of Chinese and French readers. *Journal of Cognitive Neuroscience*, 26 (3), 459-475. [10.1162/jocn_a_00499](https://doi.org/10.1162/jocn_a_00499)
- Tsigelny, I. F., Kouznetsova, V. L., Baitaluk, M. & Changeux, J.-P. (2013). A hierarchical coherent-gene-group model for brain development. *Genes, Brain and Behavior*, 12 (2), 147-165. [10.1111/gbb.12005](https://doi.org/10.1111/gbb.12005)
- Tulving, E. (1983). *Elements of episodic memory*. New York, NY: Oxford University Press.
- Uhlhaas, P. J., Roux, F., Singer, W., Haenschel, C., Sireteanu, R. & Rodriguez, E. (2009). The development of neural synchrony reflects late maturation and restructuring of functional networks in humans. *Proceedings of the National Academy of Sciences USA*, 106(24), 9866-9871. [10.1073/pnas.0900390106](https://doi.org/10.1073/pnas.0900390106)
- Waddington, C. H. (1942). The epigenotype. *Endeavour*, 18-20.
- Zhu, Y., Zhang, L., Fan, J. & Han, S. (2007). Neural basis of cultural influence on self representation. *NeuroImage*, 34 (3), 1310-1317. [10.1016/j.neuroimage.2006.08.047](https://doi.org/10.1016/j.neuroimage.2006.08.047)

Should we be Epigenetically Proactive?

A Commentary on Kathinka Evers

Stephan Schleim

“Can we be epigenetically proactive?”, is the question asked by Evers in her paper in this collection. After describing an original approach to using insights from the epigenesis of neural networks to develop new training and treatment programs, in particular to educate children and adolescents to become less violent and more sympathetic, the author suggests that there is a naturalistic responsibility for using science in this manner. In this commentary, I relate her proposal to the human enhancement debate at large, with a focus on the prevalent concept of human wellbeing. After a discussion of the factors that account for people’s quality of life and the role of research that allows them to decide the priorities for a good life themselves, three caveats against Evers’s approach are presented: (1) that epigenetic intervention carries the risk of psychological side-effects; (2) that people’s autonomy must be respected; and (3) that the world’s situation may not be as bad as suggested by the author when describing the benefits of her proposal. It is therefore concluded that, at least for the time being and until these challenges are met, we should not be epigenetically proactive.

Keywords

Adaptation | Autonomy | Neuroenhancement | Social engineering | Wellbeing

Commentator

[Stephan Schleim](#)
s.schleim@rug.nl
Rijksuniversiteit Groningen
Groningen, Netherlands

Target Author

[Kathinka Evers](#)
kathinka.evers@crb.uu.se
Uppsala Universitet
Uppsala, Sweden

Editors

[Thomas Metzinger](#)
metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Introduction

[Kathinka Evers](#) [this collection](#) discusses the possibility of changing people epigenetically. In particular, she discusses the option of increasing sympathy and decreasing xenophobia and violence. The term “*epigenetics*” is often used to describe processes affecting the activity of genes such as DNA methylation, which might enable the inheritance of acquired properties ([Bird 2007](#)). In contrast to this meaning, Evers uses the term more narrowly, with reference to the epigenesis of neural networks by selective stabilisation of synapses as an essential mechanism of

brain development ([Changeux & Danchin 1976](#)). The idea of affecting people’s development—or *ontogenesis*—through this mechanism, in order to achieve a desired state (e.g., an increase in sympathy) and/or to avoid an undesired state (e.g., a decrease in xenophobia or violence) can then be called *epigenetic proactivism*.

After describing human beings as social individualists and egocentric evaluators predisposed for selective sympathy and xenophobia, Evers explains neuronal epigenesis in detail. By influencing synaptic selection, this process may

critically affect social and cultural evolution. The central brain area for this is, according to the author, the prefrontal cortex, which is involved in planning, decision-making, thought, and socialisation; in particular, lateral prefrontal areas are associated with behaviour control. With respect to a task developed to test prefrontal cortex functioning, namely the Wisconsin Card Sorting Task (Dehaene & Changeux 1991), Evers discusses how neuronal epigenesis could explain rule-learning and top-down control. Finally, she devises two examples—adolescent violence in relation to their social environments and violence in adults associated with interconfessional conflicts—to illustrate what epigenetic proactivism may mean in practice. She eventually invokes a *naturalistic responsibility* to use the respective scientific and philosophical knowledge for the benefit of ourselves and our societies.

In this commentary, I will start out by relating Evers's proposal to the *human enhancement* debate, which has received much attention recently—in particular within neuroethics. After summarising the general assumptions and caveats of this debate, I will elaborate on the definition of people's wellbeing prevalent in the discourse on human enhancement and present an alternative based on social science research.

Finally, I will discuss epigenetic proactivism, Evers's original proposal for changing people, in more detail. Arguing that the actual means—whether neurobiological, psychological, or social—do not matter very much, while issues related to adaptation, autonomy, and instrumentalisation are of essential ethical and philosophical relevance, I will emphasise the role of an individual's *informed decision*. I will discuss in particular the three theses that (1) their proposed epigenetic intervention carries the risk of psychological side-effects; (2) that people's autonomy must be respected; and (3) that the world's situation may not be as bad as suggested by the authors when describing the benefits of their proposal. My conclusion will therefore be that the ethical justifiability of epigenetic proactivism critically depends on whether people can freely choose themselves whether or not to become epigenetically proactive, in a

situation sufficiently free from social coercion and in sufficient awareness of the likely outcomes—effects as well as side-effects—of that intervention.

2 The human enhancement debate

In a paper on the “biopolitics” of cognitive enhancement, Peter Reiner recently referred to Plato's *Phaedros*, where Socrates discusses what we nowadays might call the psychological side-effects of writing, namely the risk that our memory skills will deteriorate when we rely more on written texts (2013). Interestingly, Socrates's concerns—voiced some 2400 years ago—seem to be confirmed by recent experiments indicating that people are less likely to remember information when they expect it to be easily accessible with the aid of computers (Sparrow et al. 2011). It goes without saying that everything we do has some psychological or neural impact, whether transient or permanent. However, writing—and, more recently, digital information processing—can be seen as an enhancement technology, as it enables asynchronous and distant communication with contemporaries as well as saving thoughts and ideas for the future.

We should keep in mind, though, that the very notion of *cognitive enhancement* was introduced only recently into the scholarly debate and its increasing prevalence coincided with the institutionalisation of neuroethics in the early 2000s (Figure 1). In the meantime, some authors criticised the exaggerated promises of the debate, pointing out misperceptions in the assessment of pharmacological enhancement behaviour, the complexity of the brain's neurotransmitter systems, and the insufficient success of the much larger bio-psychiatric paradigm of improving psychological functioning in those looking for treatment (Lucke et al. 2011; Quednow 2010; Schleim 2014a). The latter means that even when the aims of the intervention are clearly circumscribed—e.g., decreasing the severity of the symptoms characteristic of a disorder—and research funds are abundant, bio-psychiatric research has unfortunately not been as successful as expected. This may relativise the hopes for effective bio-

psychological enhancement in the healthy in the near future.

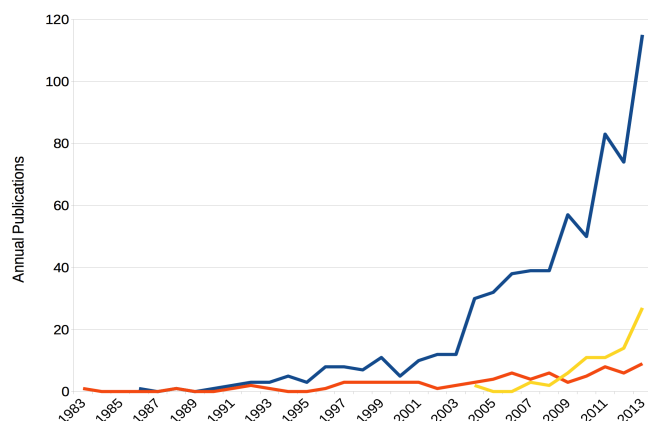


Figure 1: Publications on enhancement. Publication data from the ISI Web of Science show a steep increase in publications covering “cognitive enhancement” (blue) that coincides with the institutionalisation of neuroethics (Farah 2012). “Mood” or “affective enhancement” (orange) and “neuroenhancement” (yellow) are addressed much less frequently, although these topics also are increasingly discussed. (ISI Web of Science Topic Search)

While describing writing as a means of cognitive enhancement may seem plausible at first glance, it also carries the risk of neglecting several distinctions that may be ethically and socially important. Such distinctions are, for example, those between learning the use of an instrument to achieve a certain aim and oneself becoming an instrument for the aims of others; between using an external device and directly interfering in the body; and between defining ends autonomously and being adapted to another’s ends heteronomously. Distinctions in actual cases will not always be clear and often fall into a grey zone, but this does not mean that possible interventions cannot be discussed against these concepts. These may be understood as marking the ends of a spectrum: for example, from full autonomy to full heteronomy. Indeed, while some scholars frame the consumption of stimulus drugs such as amphetamine, methylphenidate, or modafinil by students as individual choices for better cognitive functioning (Greely et al. 2008), that is, in an autonomous fashion, several results suggest that stu-

dents might rather respond to the demands of a competitive academic environment, and thus heteronomously. I will argue later that this opposition between freedom and coercion is the crucible of ethically assessing epigenetic proactivism.

There is already empirical evidence from representative surveys or interviews with students that emphasises the relevance of this distinction. For example, M. Elizabeth Smith & Martha Farah describe in their extensive review on “smart pills” that the largest nationwide study identified admissions criteria (competitiveness) as well as two other social factors as the strongest predictors of stimulant drug consumption (2011). Interviews with non-medical consumers of stimulant drugs at an “elite” college carried out by Scott Vrecko suggest that people use stimulants for emotional and motivational ends rather than for cognitive enhancement, in particular to increase motivation to begin with or to complete boring tasks (2013). Finally, reviewing forty studies on public attitudes toward pharmacological cognitive enhancement, Kimberly J. Schelle and colleagues found that coercion to use drugs is a consistently mentioned concern (Schelle et al. 2014). This evidence associates the availability of enhancements like stimulant drugs with the pressure to adapt people to given standards of performance. Yet in the scientific literature the notion of cognitive enhancement is much more prevalent than the emotional and motivational aspects frequently mentioned in practical contexts (Figure 1).

Scientists and policy-makers in the UK *Foresight Project on Mental Capital and Well-being* note that globalisation increases demands for competitiveness as well as the pressures in our working lives (Beddington et al. 2008; Foresight Project 2008). They conclude that in a rapidly changing world like ours, we must make the most of all our resources in order to keep up with competitors; whole countries have to capitalise on their citizens’ cognitive resources. To achieve this aim, John Beddington and colleagues see vast possibilities in improving a country’s “mental capital” for all members of the population. They identify the possibility to do so at each stage in life, such as the early

identification and treatment of people with learning difficulties or the governmental support of those who want to work longer—though, notably, not shorter. A failure to react in a timely way to the challenges would come at a high cost for society, while early intervention in education could improve productivity at work and avoid costs related to a loss of mental capital (Beddington et al. 2008).

This view on performance enhancement for individual and social welfare reflects the focus of influential papers in neuroethics, emphasising the potential improvement of attention, memory, or wakefulness through the consumption of stimulant drugs or other pharmacological substances and neuroscientific technologies affecting the nervous system (Farah et al. 2004; Greely et al. 2008). Assumptions regarding the possible benefits of such substances are frequently based on trials employing test designs from *clinical psychology*, developed to identify and trace impairment in psycho-behavioural functioning, whether the investigated sample consists of patient populations, healthy people, or both (Bagot & Kaminer 2014; Repantis et al. 2010; Smith & Farah 2011).

Even if such test designs are of high clinical value, it is much less clear what statistically significant, yet often subtle, improvements in such experimental tasks, for example, in planning or memory games, mean for the *living environment* of the healthy. Whether such improvements indeed translate into an increase in individual wellbeing or the mental capital of a nation has yet to be shown. Indeed it is not even clear what a reliable and ecologically valid way of answering this question would look like. While this is still quite challenging after much debate on pharmacological enhancement, it is presently even less clear what such a standard could look like for epigenetic proactivism. In addition to measuring the benefits, neuroscientists frequently address the possibility of a psycho-behavioural trade-off—that is, the risk that an improvement in one domain would come at a loss in others (Brem et al. 2014; Hills & Hertwig 2011; Quednow 2010; Wood 2014). Given these complexities in the empirical research on

enhancement, it will be helpful to introduce an explicit definition for further discussion.

Human Enhancement =_{Df} A change in the biology or psychology of a person which increases the chances of leading a good life in the relevant set of circumstances.

Notice how this definition, proposed by Julian Savulescu and colleagues in the introduction to a recent edited volume on human enhancement (Savulescu et al. 2011), relates the good life of an individual—its biology or psychology—to the context in which that individual lives: human enhancement is something done to or with a particular person in a fixed set of circumstances, namely, a change in her or his biology or psychology. This choice already predisposes the debate and research on enhancement with respect to adapting an individual to her or his environment.

To provide an illustrative and provocative counterexample: under this definition the “treatment” of a homosexual suffering from social exclusion by instigating heterosexual acts and relations, as was routinely performed by clinical psychologists and psychiatrists until the 1970s (Barlow 1973; Hinrichsen & Katahn 1975), would qualify as a form of human enhancement—inasmuch as it succeeds in “helping” the subject to avoid the undesired sexual behaviour that instigates social exclusion and the suffering probably caused by it. With respect to this historical example we already know that leading psychiatrists later acknowledged that there was nothing inherently wrong with homosexuals, but that their suffering indeed originated from social exclusion; this reasoning eventually led to the decision not to consider homosexuality a mental disorder any longer (Friedman et al. 1976). It is instructive to contrast the definition proposed by Savulescu and colleagues with the following inverted alternative.

Human Enhancement-Inverted =_{Df} A change in the relevant set of circumstances that increases the chances of a person to lead a good life according to her or his preferences.

This alternative is not meant to be a logical inversion, but instead switches the levels of intervention, of that which is malleable and that which is considered as given. In an experimental fashion, one could also say that it is about a switch of dependent and independent variables, from the individual to its life context. Yet the aim of the intervention remains unchanged: increasing the chances of leading a good life. It goes without saying that both definitions, when put into practice, are constrained by available means and ethical principles, for example also requiring that we take the likelihood of other people's chances of leading a good life into account. It is not necessary here to argue that the inverted definition is better than the original; my intention is merely to show that we need not focus on bio-psychological changes alone. Instead, we can target the *social context* as well, decreasing the risk of adapting people to a social standard. Please note that this in itself does not imply a normative judgment, but rather widens the perspective for further analysis by taking alternative levels of intervention into consideration. As mentioned before, the balance between freedom and coercion, and autonomy and heteronomy will be essential with respect to epigenetic proactivism.

Here I have described some basic assumptions and criticism of the neuroethics debate on human enhancement, including the association of wellbeing with standards developed in clinical contexts that focus on individuals rather than on their social contexts. In the next section I will introduce research aimed at describing and understanding what people themselves consider to be quality of life, which poses an alternative to the standard adapted from clinical psychology.

3 Who defines wellbeing?

The position paper on cognitive enhancement by Henry Greely and colleagues starts out with the claim that “[s]ociety must respond to the growing demand for cognitive enhancement” (Greely et al. 2008, p. 702). The article by Beddington and colleagues on the mental wealth of nations begins with the conclusion that “[t]o

prosper and flourish in a rapidly changing world, we must make the most of all our resources—both mental and material” (Beddington et al. 2008, p. 1057). Both statements are similar in that they frame recent developments in such a way that they necessitate a reaction: we “must” respond in a particular manner. Greely and colleagues call for a “responsible use of cognitive-enhancing drugs by the healthy” (Greely et al. 2008, p. 702), though the majority of readers responding to their paper understood them as exaggerating the benefits of drug use generally or as being financially influenced by drug companies (Greely 2010). Beddington and colleagues call for the maximisation of our resources. All these authors want to increase benefits and decrease harms. However, who defines what counts as a benefit, as wellbeing, or as a good life? This is an essential and fundamental question that will influence every benefit-risk-analysis on human enhancement (Nagel 2014; Schleim 2014b).

As mentioned in the previous section, several scholars discuss the potential of means for enhancement, particularly psychopharmacological drugs, with respect to studies employing clinical test designs—whether investigating healthy people, those with a mental disorder, or even animals. Such tests measure reaction times or error rates in tasks requiring, for example, attention, memory, or planning. That is, the experimental setting frequently originates from a pragmatic context guided by identifying, treating, and/or predicting the development of a certain mental disorder. The underlying *mental disorder concept*, which is in itself controversial and subject to recurrent modifications, essentially hinges on a subject's clinically significant distress or functional impairment in the domain of cognition, emotion, and behaviour (American Psychiatric Association 2013; Stein et al. 2010). However, benefit, wellbeing, or a good life as discussed in the debate on human enhancement at large are not merely the opposites of clinically significant impairment; a five percent increase, say, in a task where a subject has to memorize as many digits as possible, and that may identify memory problems, does not reflect an increased performance in a real test, not

even a maths exam at school or university. Much less is it a suitable indicator of a benefit for the quality of life, although such a finding may be sufficient for publication in a peer-reviewed pharmacological journal.

However, there are advanced, direct, and representative measures of the quality of life. One example is the United Nations *World Happiness Report*, which compares the situation in 156 countries. The variables GDP per capita, social support, healthy life expectancy at birth, freedom to make life choices, generosity, and perceptions of corruption together explain 75.5% of the international variance of world happiness in 2012 (Helliwell et al. 2013). A more recent development is based on the OECD *Guidelines on Measuring Subjective Well-being* (OECD 2013). These allow people to create their own *Better Life Index*, prioritising eleven pre-defined domains such as education, jobs, housing, or safety.

More than 60,000 citizens from OECD countries have so far submitted their preferences, yielding important regional differences.¹ For example, people from the USA valued housing (on average 7.8 on a scale up to 10 points) and income (10.0) the highest, but work-life balance comparatively low (5.3). By contrast, people from Denmark, which is number one in the World Happiness Report, prioritised work-life balance higher than all others (9.8), and also valued life satisfaction (9.4) and community (10.0) very highly, while considering income less important (4.0). One may raise the question, of course, whether such statements are biased by social stereotypes or social desirability, but what could be a better measure of what people find important for leading a happy life than asking them directly? This is particularly so when they participate in the survey entirely on their own account.

These results emphasise two essential points for the human enhancement debate: first, people differ individually as well as regionally on what they find important for their wellbeing. Second, many of these aspects are not directly based on bio-psychological factors, but on social

factors. Indeed, the OECD construct of subjective wellbeing focuses on income, health status, social contact, employment status, personality type, and culture as determinants of life satisfaction, affect, and eudaimonic wellbeing. Unlike clinical measures of psycho-behavioural performance, they do not primarily rely on functional impairment.

Most importantly, the Better Life Index allows people to indicate themselves what they find important for their subjective wellbeing; and it turns out that many of these aspects, like housing or safety, are actual social factors that can only very indirectly be targeted by bio-psychological intervention. Therefore it becomes clear that a biased or narrowed concept of human enhancement carries the risk of missing the point of what determines or enables a better life. Further systematic analysis beyond the scope of this paper is required to show whether the factors identified are more amenable to individual psychobiological intervention, such as targeted by Savulescu and colleagues (Savulescu et al. 2011), or socio-political initiatives. Yet, while Greely and colleagues or Beddington and colleagues merely assume that increased cognitive performance will increase people's quality of life (Beddington et al. 2008; Greely et al. 2008), an initiative like the OECD Better Life Index allows people to autonomously express their own views on the issue and thus provides robust empirical evidence. This strategy helps to avoid two normative fallacies: first, that a parentalistic decision is possible when it comes to what should be good for others and, second, the idea that just because some intervention leads to a higher test score it is therefore good.

This section has highlighted, again, the tension between individual freedom and social adaptation, between autonomy and heteronomy. While most scholars would emphasise that people should be free to choose for themselves, fundamental definitions as well as the framing of human enhancement can implicitly narrow freedom, for example by introducing a limited standard for quality of life or by constraining the target for intervention. That is, when people apparently have free choice, because they are asked to choose from a number of alternatives

¹ <http://www.oecdbetterlifeindex.org>
accessed July 18, 2014

that choice may actually be quite limited, because the offered options neglect important alternatives.

As described in the previous section, people are well aware of the threat of coercion when discussing the prospects of enhancement. Coercion does not only exist at gunpoint, when acting under duress in a strong legal sense, but it can also come in a much less direct manner: For example, by telling people that they *must* choose from a limited set of options, because otherwise something bad is going to happen. Referring to what, putatively, many people are already doing or what globalisation requires increases the pressure on individuals. There are meaningful and evidence-based alternative views on human enhancement, beyond those focusing on functional impairment, as shown in this section. In the next section, I will focus on the epigenetic proactivism proposed by Kathinka Evers in more detail.

4 Epigenetic proactivism

Evers starts out their description of the naturalistic responsibility to become epigenetically proactive with a reference to the *Universal Declaration of Human Rights*. She criticises that, understood as a description of the present world, it is false to assume that all humans are born free and equal in dignity and rights; and if we understood this as a normative ideal, it would be unrealistic to guarantee these rights for every human being, given our present cerebral structure. In contrast to the human rights ideal, many people suffer from poverty and insufficient health care, and live through serious conflicts. Most people lack the sympathy necessary to respect the rights of others and all humans exhibit some kind of xenophobia. In the end, Evers even refers to the idea that humans might be subject to some built-in error or deficiency, predisposing us to self-destruction. Against this background, she proposes her epigenetic proactivism as follows:

Synaptic epigenetic theories of cultural and social imprinting on our brain architecture open the door to being epigenetically proactive, which means that we may culturally influence our brain organisation

in the aim of self-improvement, individually as well as socially and change our biological predispositions by a better fit of our brain to cultures and social structures. (Evers [this collection](#), p. 12)

She discusses two examples in more detail, namely violence in adolescents and violent interconfessional conflicts. Referring to neurodevelopmental research on children and teenagers' brains, she suggests that different educational measures such as physical exercises, cultural games, and new therapies amount to a kind of proactive epigenetic imprinting that increases control of aggression, emotion regulation, sympathy, and tolerance. It would be largely a matter of political will and social agreement, Evers claims, to develop the research enabling such educational programs and to apply them in practice. If successful, epigenetic proactivism would make societies more peaceful and inclusive, but the author also points to a problematic circularity, namely that we perhaps first need to live in an already peaceful society in order to enact such educational programs to maintain peace.

If we had to choose between epigenetic proactivism and the destruction of humankind, the decision would probably be easy; and the humbler prospect of avoiding adolescent violence and interconfessional conflicts also has some seductive allure. However, for three reasons I hesitate to agree with the conclusion that we have a naturalistic responsibility to improve ourselves epigenetically, assuming that science will develop enough at some point and offer the novel educational measures suggested by Evers: first, decreasing the disposition towards aggressive behaviour and increasing sympathy might have unexpected psychological side-effects; second, the value of human autonomy has to be considered by epigenetic proactivists, too; and third, the human condition might not be as bad as the author describes. I will discuss these three caveats in the following sections.

4.1 Side-effects of epigenetic proactivism

At first glance, who would disagree that a world with less aggression and more sympathy would

be a better world? If we could indeed decrease adolescent and interconfessional violence, why shouldn't we put such an educational program into action? Evers refers to Darwin and evolution several times in her paper. Consequently, this biological framing also raises the question of the possible evolutionary value of aggression and violence (Eibleibesfeldt 1977; Smith & Harper 1988). Darwin's original idea of the survival of the fittest emphasises the very notion of securing access to scarce resources—often at the cost of other living beings, which may even lead to the extinction of a whole species. It may well be that aggression is an essential driver of evolutionary development.

It goes without saying that from the fact that something leads to an increased survival value it does not follow that it is morally good. But it is clear that, even from a social perspective, aggression might have a function, or might be necessary for achieving some desirable ends. In the famous novel *A Clockwork Orange* by Anthony Burgess, we learn about a fictional case where a cruel and ruthless juvenile delinquent—Alex—is successfully treated bio-psychologically to stop being violent. This is carried out in a pharmacologically enhanced operant conditioning program that associates scenes of violence with aversive stimuli, such that the former delinquent feels severe nausea whenever he is confronted with aggression, including assaults against himself. This has the side effect that after the treatment Alex cannot defend himself anymore and he therefore becomes a victim of severe humiliation.

While this example is different from the case of interconfessional violence discussed by Evers, it is directly related to her other example of violence in adolescents. It is a complex bio-psychological question whether negative facets of aggression can be extinguished without also affecting people's capacity for self-defence. The author is aware of the problem of circularity, that a world may first have to become peaceful for epigenetic proactivism to be successful—and the present caveat emphasises this dilemma: if only some people were educated to avoid violence and conflicts, this could easily be abused by others.

How about increasing sympathy, then? Evers is critical about the fact that people are xenophobic and restrict their sympathy to small groups, while they should ideally extend it to human society at large. As disappointing as it may be from an ethical point of view, it could well be that a distinction between one's own or one's group's welfare from that of others is essential for psychological wellbeing. A dysfunctional self-other distinction, drawing a clear line between oneself and others, may play a role in schizophrenia (Decety & Sommerville 2003; Jardri et al. 2011). Furthermore, several investigations reported an association between emotional empathy and depression or decreased life-satisfaction (Gawronski & Privette 1997; Lee et al. 2001; O'Connor et al. 2002).

These links with mental health may be speculative to some extent, yet they illustrate that even a *prima facie* positive capacity may become negative when increased too much. Accordingly, it has become common wisdom within psychopharmacology that there is an optimal level of neurotransmitter concentration in the brain and that both a decrease and an increase may be dysfunctional and/or lead to unexpected side-effects (Wood et al. 2014). Even if ethicists, in line with Evers, presented strong arguments in favour of considering the welfare of those far away from oneself or one's group (Greene 2003; Sidgwick 1907; Singer 2002; Unger 1996), it should be born in mind that an increase of sympathy might lead to a decrease in subjective wellbeing.

4.2 Human autonomy

The vision of a scientifically enhanced world, where people are better at controlling their emotions, particularly aggression and other impulses that might lead to violent behaviour, is a recurrent topic in the history of science. For example, in the 1960s and 1970s, neuroscientists, psychologists, and sociologists all discussed the problem of delinquency and aggression, also with respect to adolescents, and proposed different solutions for coping with it. The pioneer of *brain stimulation*, José Delgado, tested the effects of electrical inhibition or excitation of dif-

ferent brain areas associated with emotion processing, such as the amygdalae, in several animal species as well as in humans (Delgado 1965, 1971; Delgado et al. 1968). His discussion of the social implications of such technology is surprisingly reminiscent of epigenetic proactivism:

Understanding of biology, physics, and other sciences facilitated the process of ecological liberation and domination. Man rebelled from natural determination and used his intelligence and skills to impose a human purpose on the development of the earth. We are now on the verge of a process of mental liberation and self-domination which is a continuation of our evolution. Its experimental approach is based on the investigation of the depth of the brain in behaving subjects. Its practical applications do not rely on direct cerebral manipulations but on the integration of neurophysiological and psychological principles leading to a more intelligent education, starting from the moment of birth and continuing throughout life, with the pre-conceived plan of escaping from the blind forces of chance and of influencing cerebral mechanisms and mental structure in order to create a future man with greater personal freedom and originality, a member of a psychocivilized society, happier, less destructive, and better balanced than present man. (Delgado 1971, p. 223; reference omitted)

He and others (e.g., Mark & Ervin 1970; Valenstein 1973) were convinced that therapeutic need would drive the development of such neurotechnology. The envisioned “psychocivilized” world would be so beneficial for individuals and society at large, Delgado believed, that the advantages overruled any social and ethical caveats (Delgado 1971). At the same time, the psychologist Burrhus Skinner wrote a best-selling book on his vision of a peaceful society realised through *social engineering* and inspired by behaviourism rather than neurotechnology (Skinner 1971). Through rewarding the right

kind of actions, Skinner suggested, the socially desired behaviour would become more likely, and the undesired behaviour more unlikely. To avoid a totalitarian regime, the people subject to this social engineering should in turn control the reward structures, the so-called contingencies of a society. Yet, in spite of the book’s popularity, it was strongly criticised by Noam Chomsky for confusing science and politics and for a misapplication of central notions such as freedom and dignity (1971).

The two utopian proposals by Delgado and Skinner, the part of the human enhancement debate discussed above that describes a need for adaptation as without alternative, and epigenetic proactivism have in common that people should be changed in such a way that they contribute to a (putatively) desired social aim: a macroscopic state with better performance, competitiveness, peacefulness, and/or caring for others. This is in obvious conflict with the notion of autonomy that is so fundamental to Immanuel Kant’s moral philosophy: no human being must be treated only as a means to another end; all humans must also be treated as an ends in themselves (1785/1994). Given the description of epigenetic proactivism by Evers, stating that our brains shall fit better to our cultures and social structures, one may well ask whether those enhanced in this manner would not become mere instruments for the present system, with its social norms and values. Also with respect to John Stuart Mill’s utilitarian liberalism, interventions to improve people seem problematic, as Mill formulated the principle:

[...] that the sole end for which mankind are warranted, individually or collectively, in interfering with the liberty of action of any of their number, is self-protection. That the only purpose for which power can be rightfully exercised over any member of a civilised community, against his will, is to prevent harm to others. His own good, either physical or moral, is not a sufficient warrant. He cannot rightfully be compelled to do or forbear because it will be better for him to do so, because it will

make him happier, because, in the opinions of others, to do so would be wise, or even right. These are good reasons for remonstrating with him, or reasoning with him, or persuading him, or entreating him, but not for compelling him, or visiting him with any evil in case he do otherwise. [...] Over himself, over his own body and mind, the individual is sovereign. (1859/1989, pp. 17–18)

Interestingly, Mill explicitly formulated the exception of self-protection and harm to others, to which Evers refers in her paper as well. However, I doubt that epigenetic proactivists can base their ethical justification on this case, as the harm they want to avoid is very indirectly related to intervention—which will most likely be applied to many people who would not have posed a threat to others without it. Furthermore, it can be doubted how imminent the danger is at all; this last point will be elaborated in the next subsection. Although other and more recent versions of “utilitarianism”, such as preference utilitarianism, place less emphasis on autonomy than Kant or Mill, they also lend the inner core of a person, for example, her or his preferences and values, a status of special protection (Singer 2011). This core is likely to be affected by changing people’s predisposition to aggression and sympathy, as the brief description of psychological side-effects in the previous subsection suggests.

Therefore, the essential question for epigenetic proactivism seems to be whether people can autonomously consent to the intervention. Evers’s title asks whether we *can* be epigenetically proactive; I have reformulated this to ask whether we *should* be epigenetically proactive. Here it is particularly relevant that her two examples, adolescent and interconfessional violence, explicitly address the development of children and teenager’s brains—that is, people whom we do not usually consider to be (fully) autonomous. The question of whether parents can take this decision, aimed at rewiring the nervous system of their children for a social aim, is too complex to be discussed here, but it calls for a solution before we can really think

about putting epigenetic proactivism into practice.

For our present purposes it shall suffice to suggest that it is unlikely that all parents would consent to such a measure. What would then happen to those who declined to participate in epigenetically proactive educational programs? Even today, some families resist education because they see a conflict between their values and teaching on, for example, sex education or evolutionary theory. In particular, those who benefit from the present social order would be unlikely to consent to a measure that might lead to a loss of power for them. As mentioned earlier, this may make those who are made less aggressive and more empathic more likely to be exploited by those who are not. Therefore, it is an essential challenge for epigenetic proactivism to take autonomy, informed consent, and the further complexities of intervening in the core of a person’s personality into account—and to consider that people’s views on these issues will be diverse!

Until these challenges of autonomy and informed consent in particular are met, I draw the tentative conclusion that we should not be epigenetically proactive. It should be noted, though, that while I am discussing the proposal by Evers here, the argument from autonomy is independent of the means actually used to enhance people—whether biological, psychological, or social. Rather, it is essential that people are free from coercion and can decide for themselves whether or not they want to become the kind of human being envisioned by proponents in the human enhancement debate, and that they have sufficient knowledge on the implications of that choice. Evers particularly focusses on children and adolescents when discussing examples of epigenetic proactivism, but it appears to be most difficult to describe what autonomous and informed choice means in precisely this group of human beings.

4.3 The human condition

Evers emphasises that many people live in precarious circumstances, even more than sixty years after the Universal Declaration of Human

Rights; in the end, she even refers to Arthur Koestler's idea that humans might have some built-in deficiency, predisposing us to self-destruction. Obviously, against that prospect, the promises of epigenetic proactivism look seductive. Indeed, we must concede that even some twenty-five years after the Cold War international conflicts have not abated altogether—in some areas they have even multiplied, and terrorism or economic instability are a concern for many. However, from the perspective of cultural evolution, universal human rights are a rather novel development and it may be too early to take a pessimistic stance on their success and effect. Returning to the UN World Happiness Report (Helliwell et al. 2013), one may ask whether the difference between the leading countries—Denmark, Norway, Switzerland, the Netherlands, and Sweden (ranked 1st to 5th)—, those in the middle—Libya, Bahrain, Montenegro, Pakistan, and Nigeria (ranked 78th to 82nd)—, and those at the bottom—Rwanda, Burundi, the Central African Republic, Benin, and Togo (ranked 152nd to 156th)—can be explained or even overcome by means of human enhancement like epigenetic proactivism rather than internationally-aided institutional development.

One shared rhetorical feature of those visions of a better humankind is a claim that all has somehow gone wrong, and even to predict an imminent catastrophe. For example, the various *Humanist Manifestos* of the 20th and early 21st century described serious threats to human survival.² Delgado emphasized an imbalance between our material and mental evolution, putting humanity at risk (1971), and Skinner started out by referring to problems related to population growth, pollution of the environment, and nuclear armament (1971). It probably lies in the eye of the beholder to speculate whether humankind has not yet destroyed itself because or in spite of unprecedented technological powers.

It is a matter of fact that we have not yet done so, and although many things have gone

wrong, others have gone right. Steven Pinker recently gathered evidence that, particularly when viewed in relation to the vast population growth of humanity, our present times are much more peaceful than the past (2011). He describes processes of pacification and civilization as well as a humanitarian and rights revolution that can provide hope that things will change for the better, not only for the worse. Therefore, even if human enhancement in general or epigenetic proactivism in particular may offer genuine improvement of the human condition in several ways, they are probably not necessary for human survival.

5 Conclusion

Kathinka Evers summarises research on the epigenesis of neural networks to describe a vision of epigenetical proactivism, a development of new training and therapeutic programs to improve humans. She asks whether we *can* be epigenetically proactive, pointing out the benefits of decreasing the prevalence of adolescent and interconfessional violence, and in so doing develops her answer: yes, in principle, we can be epigenetically proactive. However, she also describes a naturalistic responsibility to do this, which is the point at which my discussion of her proposal diverged from her view. Particularly with respect to autonomy and free choice I think that, for the time being, we should not be epigenetically proactive; and we should be even more cautious when interventions in children's and teenagers' brains are at issue. Minor caveats are related to the possible psychological side-effects of decreasing our disposition towards aggression and increasing that of sympathy, as well as a more optimistic view of how humankind is developing.

In this paper, I also related epigenetic proactivism to the human enhancement debate more generally, which has become much more comprehensive than can be addressed in such a brief commentary. It was important to examine the definition of wellbeing and the framing of urgency, as well as the primary level of intervention—bio-psychological or social—, issues that are also related to autonomy. This does not

² See the three Manifestos of 1933, 1973, and 2003 of the American Humanist Association on <http://americanhumanist.org/Humanism/> (accessed July 21 2014).

mean that knowledge on epigenetics could not be used in another manner for the purposes of enhancement, in situations where people can make an informed decision for themselves whether and how to engage in a certain kind of training. In this sense, it would be interesting to compare epigenetic proactivism to other non-pharmaceutical means of enhancement, such as nutrition, exercise, sleep, or meditation (Dresler et al. 2013). Generally speaking, the knowledge described by Evers could also be related to debates on improving school education neuroscientifically (Hook & Farah 2013; Posner & Rothbart 2005). Furthermore, when targeting human capacities that are also salient for moral cognition, the debate on *moral enhancement* may be an important reference point with overlapping prospects and concerns (Douglas 2008, 2013; Harris 2011).

Evers warned that science has been hijacked repeatedly throughout history and that in particular the dream of creating perfect human beings has a sordid past. Here I wholeheartedly agree with her and her related call for historic awareness. I hope that I have succeeded in showing why, beyond this awareness, it is also essential to take people's own views and autonomy into account. It may not only be the case that too much focus on enhancing people makes them sad by focusing too much on their deficiencies (Schleim 2014b; Schopenhauer 1874), but in the attempt to create superhuman beings a human catastrophe might also be provoked.

Acknowledgments

I would like to thank the two editors as well as two anonymous reviewers for their extraordinarily helpful and constructive comments on a previous version of this paper.

References

- American Psychiatric Association, (Ed.) (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.
- Bagot, K. S. & Kaminer, Y. (2014). Efficacy of stimulants for cognitive enhancement in non-attention deficit hyperactivity disorder youth: A systematic review. *Addiction*, 109 (4), 547-557. [10.1111/add.12460](https://doi.org/10.1111/add.12460)
- Barlow, D. H. (1973). Increasing heterosexual responsiveness in the treatment of sexual deviation: A review of the clinical and experimental evidence. *Behavior Therapy*, 4 (5), 655-671. [10.1016/s0005-7894\(73\)80158-3](https://doi.org/10.1016/s0005-7894(73)80158-3)
- Beddington, J., Cooper, C. L., Field, J., Goswami, U., Huppert, F. A., Jenkins, R. & Thomas, S. M. (2008). The mental wealth of nations. *Nature*, 455 (7216), 1057-1060. [10.1038/4551057a](https://doi.org/10.1038/4551057a)
- Bird, A. (2007). Perceptions of epigenetics. *Nature*, 447 (7143), 396-398. [10.1038/nature05913](https://doi.org/10.1038/nature05913)
- Brem, A. K., Fried, P. J., Horvath, J. C., Robertson, E. M. & Pascual-Leone, A. (2014). Is neuroenhancement by noninvasive brain stimulation a net zero-sum proposition? *NeuroImage*, 85 (3), 1058-1068. [10.1016/j.neuroimage.2013.07.038](https://doi.org/10.1016/j.neuroimage.2013.07.038)
- Changeux, J. P. & Danchin, A. (1976). Selective stabilization of developing synapses as a mechanism for specification of neuronal networks. *Nature*, 264 (5588), 705-712. [10.1038/264705a0](https://doi.org/10.1038/264705a0)
- Chomsky, N. (1971). The case against B.F. Skinner. *The New York Review of Books*, 17 (11), 18-24.
- Decety, J. & Sommerville, J. A. (2003). Shared representations between self and other: A social cognitive neuroscience view. *Trends in Cognitive Sciences*, 7 (12), 527-533. [10.1016/j.tics.2003.10.004](https://doi.org/10.1016/j.tics.2003.10.004)
- Dehaene, S. & Changeux, J. P. (1991). The wisconsin card sorting test: Theoretical analysis and modeling in a neuronal network. *Cerebral Cortex*, 1 (1), 62-79. [10.1093/cercor/1.1.62](https://doi.org/10.1093/cercor/1.1.62)
- Delgado, J. M. (1965). Sequential behavior induced repeatedly by stimulation of the red nucleus in free monkeys. *Science*, 148 (3675), 1361-1363. [10.1126/science.148.3675.1361](https://doi.org/10.1126/science.148.3675.1361)
- (1971). *Physical control of the mind; Toward a psychocivilized society*. New York, NY: Harper & Row.
- Delgado, J. M., Mark, V., Sweet, W., Ervin, F., Weiss, G., Bach, Y. R. G. & Hagiwara, R. (1968). Intracerebral radio stimulation and recording in completely free patients. *Journal of Nervous and Mental Disease*, 147 (4), 329-340.

- Douglas, T. (2008). Moral enhancement. *Journal of Applied Philosophy*, 25 (3), 228-245. [10.1111/j.1468-5930.2008.00412.x](https://doi.org/10.1111/j.1468-5930.2008.00412.x)
- (2013). Moral enhancement via direct emotion modulation: A reply to John Harris. *Bioethics*, 27 (3), 160-168. [10.1111/j.1467-8519.2011.01919.x](https://doi.org/10.1111/j.1467-8519.2011.01919.x)
- Dresler, M., Sandberg, A., Ohla, K., Bubltz, C., Trenado, C., Mroczko-Wasowicz, A. & Repantis, D. (2013). Non-pharmacological cognitive enhancement. *Neuropharmacology*, 64, 529-543. [10.1016/j.neuropharm.2012.07.002](https://doi.org/10.1016/j.neuropharm.2012.07.002)
- Eibleibesfeldt, I. (1977). Evolution of destructive aggression. *Aggressive Behavior*, 3 (2), 127-144. [10.1002/1098-2337\(1977\)3:2<127::AID-AB2480030204>3.0.CO;2-Y](https://doi.org/10.1002/1098-2337(1977)3:2<127::AID-AB2480030204>3.0.CO;2-Y)
- Evers, K. (2015). Can we be epigenetically proactive? In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Farah, M. J. (2012). Neuroethics: The ethical, legal, and societal impact of neuroscience. *Annual Review of Psychology*, 63, 571-591. [10.1146/annurev.psych.093008.100438](https://doi.org/10.1146/annurev.psych.093008.100438)
- Farah, M. J., Illes, J., Cook-Deegan, R., Gardner, H., Kandel, E., King, P. & Wolpe, P. R. (2004). Neurocognitive enhancement: What can we do and what should we do? *Nature Reviews Neuroscience*, 5 (5), 421-425. [10.1038/nrn1390](https://doi.org/10.1038/nrn1390)
- Foresight Project, (2008). *Final project report*. London, UK: The Government Office for Science.
- Friedman, R. C., Green, R. & Spitzer, R. L. (1976). Reassessment of homosexuality and transsexualism. *Annual Review of Medicine*, 27, 57-62. [10.1146/annurev.me.27.020176.000421](https://doi.org/10.1146/annurev.me.27.020176.000421)
- Gawronski, I. & Privette, G. (1997). Empathy and reactive depression. *Psychological Reports*, 80 (3), 1043-1049. [10.2466/pr0.1997.80.3.1043](https://doi.org/10.2466/pr0.1997.80.3.1043)
- Greely, H. (2010). Enhancing brains: What are we afraid of? *Cerebrum*, 14, 1-10.
- Greely, H., Sahakian, B., Harris, J., Kessler, R. C., Gazzaniga, M., Campbell, P. & Farah, M. J. (2008). Towards responsible use of cognitive-enhancing drugs by the healthy. *Nature*, 456 (7223), 702-705. [10.1038/456702a](https://doi.org/10.1038/456702a)
- Greene, J. (2003). From neural “is” to moral “ought”: What are the moral implications of neuroscientific moral psychology? *Nature Reviews Neuroscience*, 4 (10), 846-849. [10.1038/nrn1224](https://doi.org/10.1038/nrn1224)
- Harris, J. (2011). Moral enhancement and freedom. *Bioethics*, 25 (2), 102-111. [10.1111/j.1467-8519.2010.01854.x](https://doi.org/10.1111/j.1467-8519.2010.01854.x)
- Helliwell, J., Layard, R. & Sachs, J. (Eds.) (2013). *World happiness report 2013*. New York, NY: Sustainable Development Solutions Network, a Global Initiative for the United Nations.
- Hills, T. & Hertwig, R. (2011). Why aren't we smarter already: Evolutionary trade-offs and cognitive enhancements. *Current Directions in Psychological Science*, 20 (6), 373-377. [10.1177/0963721411418300](https://doi.org/10.1177/0963721411418300)
- Hinrichsen, J. J. & Katahn, M. (1975). Recent trends and new developments in the treatment of homosexuality. *Psychotherapy-Theory Research and Practice*, 12 (1), 83-92. [10.1037/h0086413](https://doi.org/10.1037/h0086413)
- Hook, C. J. & Farah, M. J. (2013). Neuroscience for educators: What are they seeking, and what are they finding? *Neuroethics*, 6 (2), 331-341. [10.1007/s12152-012-9159-3](https://doi.org/10.1007/s12152-012-9159-3)
- Jardri, R., Pins, D., Lafargue, G., Very, E., Ameller, A., Delmaire, C. & Thomas, P. (2011). Increased overlap between the brain areas involved in self-other distinction in schizophrenia. *PLoS One*, 6 (3), e17500. [10.1371/journal.pone.0017500](https://doi.org/10.1371/journal.pone.0017500)
- Kant, I. (1994). *Grundlegung zur Metaphysik der Sitten*. Hamburg, GER: Meiner.
- Lee, H. S., Brennan, P. F. & Daly, B. J. (2001). Relationship of empathy to appraisal, depression, life satisfaction, and physical health in informal caregivers of older adults. *Research in Nursing & Health*, 24 (1), 44-56. [10.1002/1098-240x\(200102\)24](https://doi.org/10.1002/1098-240x(200102)24)
- Lucke, J. C., Bell, S., Partridge, B. & Hall, W. D. (2011). Deflating the neuroenhancement bubble. *American Journal of Bioethics Neuroscience*, 2 (4), 38-43. [10.1080/21507740.2011.611122](https://doi.org/10.1080/21507740.2011.611122)
- Mark, V. H. & Ervin, F. R. (1970). *Violence and the brain*. New York, NY: Harper & Row.
- Mill, J. S. (1989). *On liberty*. London, UK: The Walter Scott Publishing Co.
- Nagel, S. K. (2014). Enhancement for well-being is still ethically challenging. *Frontiers in Systems Neuroscience*, 8 (72). [10.3389/fnsys.2014.00072](https://doi.org/10.3389/fnsys.2014.00072)
- O'Connor, L. E., Berry, J. W., Weiss, J. & Gilbert, P. (2002). Guilt, fear, submission, and empathy in depression. *Journal of Affective Disorders*, 71 (1-3), 19-27. [10.1016/s0165-0327\(01\)00408-6](https://doi.org/10.1016/s0165-0327(01)00408-6)
- OECD, (2013). *OECD guidelines on measuring subjective well-being*. Paris, FR: OECD Publishing.
- Pinker, S. (2011). *The better angels of our nature: Why violence has declined*. New York, NY: Viking.
- Posner, M. I. & Rothbart, M. K. (2005). Influencing brain networks: Implications for education. *Trends in Cognitive Sciences*, 9 (3), 99-103. [10.1016/j.tics.2005.01.007](https://doi.org/10.1016/j.tics.2005.01.007)

- Quednow, B. B. (2010). Ethics of neuroenhancement: A phantom debate. *BioSocieties*, 5 (1), 153-156. [10.1057/biosoc.2009.13](https://doi.org/10.1057/biosoc.2009.13)
- Reiner, P. B. (2013). Biopolitics of cognitive enhancement. In E. Hildt & A. Franke (Eds.) *Cognitive enhancement: An interdisciplinary perspective* (pp. 189-200). Dordrecht, NL: Springer.
- Repantis, D., Schlattmann, P., Laisney, O. & Heuser, I. (2010). Modafinil and methylphenidate for neuroenhancement in healthy individuals: A systematic review. *Pharmacological Research*, 62 (3), 187-206. [10.1016/j.phrs.2010.04.002](https://doi.org/10.1016/j.phrs.2010.04.002)
- Savulescu, J., Sandberg, A. & Kahane, G. (2011). Well-being and enhancement. In J. Savulescu, R. H. J. ter Meulen & Kahane (Eds.) *Enhancing human capacities* (pp. 3-18). Oxford, UK: Wiley-Blackwell.
- Schelle, K. J., Faulmüller, N., Caviola, L. & Hewstone, M. (2014). Attitudes towards pharmacological cognitive enhancement – A review. *Frontiers in Systems Neuroscience*, 8 (53). [10.3389/fnsys.2014.00053](https://doi.org/10.3389/fnsys.2014.00053)
- Schleim, S. (2014a). Critical neuroscience – or critical science? A perspective on the perceived normative significance of neuroscience. *Frontiers in Human Neuroscience*, 8 (336). [10.3389/fnhum.2014.00336](https://doi.org/10.3389/fnhum.2014.00336)
- (2014b). Whose well-being? Common conceptions and misconceptions in the enhancement debate. *Frontiers in Systems Neuroscience*, 8 (148). [10.3389/fnsys.2014.00148](https://doi.org/10.3389/fnsys.2014.00148)
- Schopenhauer, A. (1874). *Parerga und Paralipomena, Band I*. Zurich, SUI: Haffmans.
- Sidgwick, H. (1907). *The methods of ethics*. New York, NY: Macmillan and Co.
- Singer, P. (2002). *One world : The ethics of globalization*. New Haven, CT: Yale University Press.
- (2011). *Practical ethics*. Cambridge, UK: Cambridge University Press.
- Skinner, B. F. (1971). *Beyond freedom and dignity*. Toronto, Canada: Bantam.
- Smith, M. E. & Farah, M. J. (2011). Are prescription stimulants “smart pills”? The epidemiology and cognitive neuroscience of prescription stimulant use by normal healthy individuals. *Psychological Bulletin*, 137 (5), 717-741. [10.1037/a0023825](https://doi.org/10.1037/a0023825)
- Smith, J. M. & Harper, D. G. C. (1988). The evolution of aggression: Can selection generate variability? *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*, 319 (1196), 557-570. [10.1098/rstb.1988.0065](https://doi.org/10.1098/rstb.1988.0065)
- Sparrow, B., Liu, J. & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, 333 (6043), 776-778. [10.1126/science.1207745](https://doi.org/10.1126/science.1207745)
- Stein, D. J., Phillips, K. A., Bolton, D., Fulford, K. W. M., Sadler, J. Z. & Kendler, K. S. (2010). What is a mental/psychiatric disorder? From DSM-IV to DSM-V. *Psychological Medicine*, 1759 (1765), 1759-1765. [10.1017/s0033291709992261](https://doi.org/10.1017/s0033291709992261)
- Unger, P. K. (1996). *Living high and letting die: Our illusion of innocence*. New York, UK: Oxford University Press.
- Valenstein, E. S. (1973). *Brain control*. New York, NY: Wiley.
- Vrecko, S. (2013). Just how cognitive is “Cognitive Enhancement”? On the significance of emotions in university students’ experiences with study drugs. *American Journal of Bioethics Neuroscience*, 4 (1), 4-12. [10.1080/21507740.2012.740141](https://doi.org/10.1080/21507740.2012.740141)
- Wood, S., Sage, J. R., Shuman, T. & Anagnostaras, S. G. (2014). Psychostimulants and cognition: A continuum of behavioral and cognitive activation. *Pharmacological Reviews*, 66 (1), 193-221. [10.1124/pr.112.007054](https://doi.org/10.1124/pr.112.007054)

Understanding Epigenetic Proaction

A Reply to Stephan Schleim

Kathinka Evers

Epigenetic proaction can be described as a way of steering evolution by influencing the cultural imprints stored in our brains. It is not to be confused with “human enhancement”. It is a process on the societal level that need not conflict with the notion of autonomy, nor suggest any “superhuman” ideal. Risks of misuse justify precaution, not abandonment of constructive scientific pursuits. Scientific knowledge can help us improve our life conditions in the long-term. A naturalistic responsibility is born out of science’s strong social relevance.

Keywords

Autonomy | Enhancement | Epigenetic proaction | Precaution | Responsibility

Author

[Kathinka Evers](#)

kathinka.evers@crb.uu.se

Uppsala Universitet
Uppsala, Sweden

Commentator

[Stephan Schleim](#)

s.schleim@rug.nl

Rijksuniversiteit Groningen
Groningen, Netherlands

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

Epigenetic proaction can be described as a way of steering evolution by influencing the cultural imprints that are stored in our brains. The question analysed in my target article is what exactly this means and whether it is possible. Can we adapt our societies to constructively interact with the ever-developing neuronal architecture of our brains? The issue of whether such interaction is desirable is also raised but not discussed in depth.

In order to decide whether an action should be pursued it would be wise to first attempt to understand its nature and implications. Regrettably, in his commentary to my article, Stephan Schleim fails to acknowledge the main concern of my paper, namely the sci-

entific issue, moving instead to the normative question via some less relevant detours. The commentary therefore becomes misleading. Rather than engaging with the scientific points I make, Schleim takes as a starting point a flawed understanding of epigenetic proaction and tries to show how undesirable it would be. The arguments have little to do with the article on which he purports to comment.

2 Confusing epigenetic proaction with human enhancement

After making the assertion that “the actual means—whether neurobiological, psychological,

or social—do not matter very much” in his philosophical analysis of epigenetic proaction, [Schleim](#) proceeds to relate my position to the general debate on “human enhancement” ([this collection](#), p. 2). A long discussion follows about this debate that, although quite popular amongst some contemporary philosophers, is here out of context. In the target article, there is no mention of individual cognitive, moral, or performance enhancement, nor any mention of pharmaceutical “smart pills” and so on. The target article does not speak of epigenetic proaction as an individual opt-in/opt-out thing at all, nor does it speak of enhancement. And it certainly does not recommend, as [Schleim](#) suggests at the end of his commentary “the attempt to create superhuman beings” ([this collection](#), p. 15). The statement that my theory proposes methods for parents “aimed at rewiring the nervous system of their children for a social aim” ([Schleim this collection](#), p. 10) is a caricature. Perhaps the author has not read the target paper quite thoroughly enough. This would explain why the author does not specifically address any of the scientific issues raised in the paper.

3 Well-being and exaggerated virtues

In the commentary, the subsequent discussion is about who defines well-being and how. Whilst this in itself is an interesting question that deserves careful consideration from many perspectives, it is not directly relevant to the target article. The article raises the question of whether epigenetic proaction is possible, and presents scientific data and theories to explain what this means. On that basis, I suggest that they can be taken to support the view that it may indeed be possible. The questions of defining well-being or of specifying who should be in charge of defining well-being, whilst interesting, fall out of this scope.

In contrast, the question of “side-effects” can with some effort be considered at least somewhat relevant to the article under debate. Here, [Schleim](#) wonders: is it possible, e.g., to reduce aggression without making a person weak or meek? Can a less aggressive person defend

him- or herself against a more aggressive person? He seems to be doubtful, but my short reply is: obviously, yes. Much education, of children in particular and in human societies in general, includes attempts to check aggression—it does not thereby create either wimps or zombies. Even martial arts focus explicitly on checking aggression, whilst by definition aiming to make students excellent in combat. [Schleim](#) also wonders about the risky side-effects of increasing sympathy. He warns that increasing sympathy too much could perhaps lead to a “dysfunctional self–other distinction” that “may play a role in schizophrenia”. However, even if this were the case, this is not a necessary—or even very common—side-effect of increasing sympathy. Certainly, when we bring our children up to sympathise with others, we may increase their distress at the sight of suffering in others, but I do not believe that we thereby increase their risk of developing schizophrenia. Moreover, as a general principle, that an initially positive value can become negative if exaggerated does not entail that we should stop seeking it altogether. If that were the case, we would have little to strive for.

4 Epigenetic proaction: A process on the societal level

[Schleim](#) compares my theory to the famously misconceived social engineering projects of Skinner and Delgado, for whom, [Schleim](#) says, the goals blessed the means. He argues ([Schleim this collection](#), p. 9) that these “utopian proposals” stand “in obvious conflict with the notion of autonomy”, as understood by Immanuel Kant: no being must be treated only as a means to an end, but as an end in itself. I agree with Kant’s principle and see no conflict between it and the notion of epigenetic proaction. There is nothing in the idea of epigenetic proaction as I develop it in my article that suggests treating people as mere means to a social end, or of allowing them to “become mere instruments for the present system” ([Schleim this collection](#), p. 9). The idea in itself is neutral in this regard: of course the idea can be misused—all science can be misused—but it is no part of the theory to

have this negative consequence. In other words, there is no essential conflict between human autonomy and human epigenetic proaction properly understood.

As for the issue of informed consent that Schleim raises in that context, it does not directly arise through the topics I address in my article, but it would arise in the research that I recommend be pursued. Epigenetic proaction is a process on the societal level. When, for example, educational structures and methods are adopted in a functioning democratic society, people are invited to express their views through political elections, public debates, consensus conferences, etc.; but we do not ask each citizen for an individual informed consent. Nor do we ask for it when laws are passed. For example, in 1979, corporeal punishment of children became illegal in Sweden. The decision was preceded (and followed) by public debate and, as with most rules and regulations, some agreed with the ruling, while others did not—but the question of informed consent does not here arise. In contrast, if research in the natural and social sciences collaborate, e.g., to develop educational structures to assist and protect adolescents during that difficult phase of cerebral development, insofar as such research requires the use of human subjects individual informed consent will be needed. That this is the case is not a specific problem of the theory, but an ethical regulation (amongst many others) that all research must respect.

5 Opposing world-views

Concerning the human condition, surprisingly, Schleim criticises me for being overly concerned about the present states of poverty, war, and the many current violations of human rights around the world. He dismisses these worries as “rhetorical” (again comparing my arguments to those of Skinner and Delgado). Schleim seems to be at relative ease with the present state and future of humanity and, referencing Steven Pinker, draws the conclusion that there is hope that things will change for the better, so there is no need to be epigenetically proactive. Different world-views here confront one another.

Schleim concludes in what seems to me again a spirit of denial that people might be saddened by “focusing too much on their deficiencies” and ends his commentary by saying that “in the attempt to create superhuman beings a human catastrophe might also be provoked” ([Schleim this collection](#), p. 12). True, no doubt—as, notably, Germany’s recent past illustrates. But this is not particularly relevant to my article: there is nothing in the theory of epigenetic proaction to suggest that we either should or could create “superhumans”.

6 Conclusion

Trying to understand and influence human norms in the light of what we today know about the brain is not an easy task. The scientific challenge is increased by the remarkable emotionality with which this whole area of research is permeated and which can apparently make it hard to see clearly what is actually being said. This emotionality is in part understandable: the notion of improving the human condition, including our biology, comes in some very sordid versions, as ideas of “racial purity” or “ethnic supremacy” serve to illustrate, and which remain present in various societies around the world. Historic awareness is indeed essential to safeguard constructive and hope-inspiring scientific ideas from being hijacked by nefarious ideologies (or, indeed, interpretations) and abused for unscientific purposes. However, the risk of misuse justifies precaution, not abandonment of constructive scientific pursuits.

Research collaborations between neuroscience, genetics and social science, notably, today provide rich and multifaceted knowledge about the human being and an increasingly integrated view of us as biological organisms interacting in complex natural and cultural environments in constant evolution. The resulting knowledge could further help us improve our life conditions, e.g., by assisting us in finding remedies for the developmental crises of adolescents, or excessive societal violence. What I call our “naturalistic responsibility” is born out of science’s strong social relevance. Whether or not in the future we shall use this knowledge

soundly remains to be seen. Which traits we decide to favour epigenetically, or what social structures we choose to develop, depends on who “we” are, and on the society in which we wish to live. We may hope that young scientists and philosophers shall rise well to that challenge, and develop the idea of epigenetic proactivity into a dynamic and socially responsible area of research.

References

Schleim, S. (2015). Should we be epigenetically proactive? A commentary on Kathinka Evers. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.

The Paradigmatic Body

Embodied Simulation, Intersubjectivity, the Bodily Self, and Language

Vittorio Gallese & Valentina Cuccio

In this paper we propose a way in which cognitive neuroscience could provide new insights on three aspects of social cognition: intersubjectivity, the human self, and language. We emphasize the crucial role of the body, conceived as the constitutive source of pre-reflective consciousness of the self and of the other. We provide a critical view of contemporary social cognitive neuroscience, arguing that the brain level of description is a necessary but not sufficient condition for studying intersubjectivity, the human self, and language; which are only properly visible if coupled with a full appreciation of their intertwined relationship with the body. We introduce mirror mechanisms and embodied simulation and discuss their relevance to a new account of intersubjectivity and the human self. In this context, we focus on a specifically human modality of intersubjectivity: language. Aspects of social cognition related to language are discussed in terms of embodiment, while emphasizing the progress and limitations of this approach. We argue that a key aspect of human language consists in its decoupling from its usual denotative role, hence manifesting its power of abstraction. We discuss these features of human language as instantiations of the Greek notion of *paradeigma*, originally explored by Aristotle to refer to a typical form of rhetorical reasoning and relate it to embodied simulation. Paradigmatic knowledge connects the particular with the particular, moving from the contingent particular situation to an exemplary case. Similarly, embodied simulation is the suspension of the “concrete” application of a process: reuse of motor knowledge in the absence of the movement it realizes is an example of “paradigmatic knowledge.” This new epistemological approach to intersubjectivity generates predictions about the intrinsic functional nature of our social cognitive operations, cutting across, and not subordinated to, a specific ontology of mind.

Keywords

Cognitive neuroscience | Embodied simulation | Intersubjectivity | Language | Mirror neurons | Paradigm | Social cognition

1 Introduction

The last decades of the twentieth century were marked by great progress in cognitive neuroscience, made possible by recently-developed brain imaging technologies such as functional magnetic resonance imaging (fMRI)—which allowed for the first time non-invasive study of the human brain.

But what is cognitive neuroscience? We think it is fair to say that it is above all a methodological approach whose results are strongly

influenced by which questions are being asked and how. Studying single neurons and/or the brain does not necessarily predetermine the questions to be asked that will help us understand how and how much our human nature depends upon our brains. Even less so the answers. Our purpose here is twofold. On the one hand, we aim to provide a brief overview of current cognitive neuroscience and its methods. We first present the limitations displayed by most

Authors

[Vittorio Gallese](#)

vittorio.gallese@unipr.it

Università degli Studi di Parma
Parma, Italy

[Valentina Cuccio](#)

valentina.cuccio@unipa.it

Università degli Studi di Palermo
Palermo, Italy

Commentator

[Christian Pfeiffer](#)

christian.pfeiffer@epfl.ch

Ecole Polytechnique Fédérale
Lausanne, Switzerland

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

current mainstream cognitive neuroscience, followed by a proposal for an alternative approach, both in terms of the employed methodology and of its main goals. In short, in contrast with what many normally take for granted¹, we assume the brain level of description to be a necessary but not sufficient condition for studying intersubjectivity, language, and the human self, which are only properly visible if coupled with a full appreciation of their intertwined relationship with the body. This overview has been specifically designed to provide a useful tool for researchers working in the humanities. Section 2 is entirely devoted to this goal.

On the other hand, the authors of this essay are a cognitive neuroscientist and a philosopher of language, and as such our second purpose is to propose how cognitive neuroscience could provide new insights on specific aspects of human cognition. In section 3 we introduce mirror mechanisms and embodied simulation and, in the following sections, we discuss their relevance for a new account of intersubjectivity, the human self, and language—which privileges the body as the transcendental foundation of each.

We emphasize the crucial role of the body, conceived as the constitutive source of pre-reflective consciousness of the self and of the other and as the ground upon which linguistic meaning is also based. The body we talk about in this paper manifests itself in two different, complementary, and closely intertwined ways:² it is a *Leib*, a lived body entertaining experiences of self and others, and a *Körper*, the somatic object, of which the brain is a con-

stitutive part. We posit that this dual nature of our experienced body can be fully understood—and its genesis revealed—by investigating its motor neurophysiological underpinnings at the sub-personal level. The naturalization of intersubjectivity, the self, and language implies a first attempt to isolate the constituent components of the concepts we use to refer to these aspects of human social cognition by literally investigating what they are made of at the level of description of the brain–body system. This attempt in relation to the notions of intersubjectivity, the self, and language is intended here to form an identification of their constitutive mechanisms. We believe that this investigation becomes really effective only when it is framed within both comparative and developmental perspectives.³

The comparative perspective not only allows us to frame human social cognition within an evolutionary picture, thus providing access to its phylogenetic antecedents.⁴ It also greatly reduces the risk of the empirical investigation of the human brain being subordinated to a specific human ontology of mind. A further reason for privileging the comparative perspective resides in the fact that it also brings us the most finely grained approach to date for studying the brain, and the possibility of correlating single neurons' activity with behaviour and cognition—as when studying single neurons' activity in non-human primates, like macaque monkeys.

The immanent transcendence⁵ of the body's corporeality can be revealed, we contend, by bringing the analysis back to the level of the brain–body system; that is, to the level of the *Körper*. We show that this particular neurocog-

¹ “Still other accounts of grounded cognition focus on situated action, social interaction, and the environment (e.g., Barsalou 2003, Barsalou et al. 2007a, Glenberg 1997, W. Prinz 1997, Rizzolatti & Craighero 2004, Robbins & Aydede 2007, E. Smith & Semin 2004, Yeh & Barsalou 2006). From this perspective, the cognitive system evolved to support action in specific situations, including social interaction. These accounts stress interactions between perception, action, the body, the environment, and other agents, typically during goal achievement. It is important to note that the phrase ‘embodied cognition’ is often used when referring to this collection of literatures. Problematically, however, ‘embodied cognition’ produces the mistaken assumption that all researchers in this community believe that bodily states are necessary for cognition and that these researchers focus exclusively on bodily states in their investigations. Clearly, however, cognition often proceeds independently of the body, and many researchers address other forms of grounding.” (Barsalou 2008, p. 619)

² We find it useful to employ here the distinction originally proposed by Edmund Husserl in Husserl (1973, p. 119).

³ One of the major contributions to our understanding of human social cognition is provided by developmental psychology. In this paper, for sake of concision we don't focus on developmental aspects, in spite of the crucial importance we attribute to them to thoroughly address the issues we want to address here.

⁴ It is interesting to note that this comparative perspective seems to support the hypothesis of an evolutionary continuity between human and non-human primates in relation to the emergence of language from our sensori-motor abilities. For a discussion of this topic, see Glenberg & Gallese (2012).

⁵ The expression immanent transcendence is meant to signify here the fact that the body, by means of a biological mechanism (immanent), can transcend his usual function (for example, motion) to become expression (a model or paradigma) of this bodily knowledge.

nitive approach is beginning to reveal the tight relationship between a core notion of the bodily self, its potentiality for action, and motor simulation at the level of the cortical motor system. Cognitive neuroscience can enable the analysis of several concepts and notions we normally refer to when describing ourselves and our social cognitive lives. In the present paper we apply this method to the notions of intersubjectivity, the self, and language.

To fully account for the specific quality of human social cognition one cannot undervalue the linguistic dimension. For this reason, we introduce aspects of social cognition related to language and discuss them in terms of embodiment, emphasizing the progress and limitations of this approach. Traditionally, the linguistic and corporeal sensorimotor dimensions of social cognition have been considered entirely unrelated. We posit that embodied simulation, conceived of as a model for important aspects of our relation to the world, might help in overcoming this apparently unsolvable dichotomy. We argue that a key aspect defining the unique specificity of human language consists in its decoupling from its usual denotative role. This means that language allows us to talk about general concepts such as beauty or mankind, without denoting any particular instance of these concepts. In so doing, human language manifests its power for abstraction. We discuss these features of human language as instantiations of the Greek notion of *paradeigma*, originally explored by Aristotle, and relate it to embodied simulation. When a word or syntagm, like the Latin word *Rosa*, is decoupled from its usual denotative role, it can function as a general rule of knowledge, e.g., as a paradigm for the female nominative case of Latin nouns belonging to the first declension. The notion of *paradeigma* does not establish a connection between a universal principle and its contingent aspects, as in deduction, it rather exemplifies a particular case of induction: specifically the transposition of inductive reasoning in the field of studies of persuasive communication, known as rethorics, where contingent particular cases lead to general rules describing them. Paradigmatic knowledge, however, differently

from standard cases of induction, connects the particular with the particular, moving from the contingent particular situation to an exemplary case. We propose that embodied simulation could instantiate such a notion of paradigmatic knowledge, hence enabling its naturalization and helping us overcome the apparent gap between the linguistic and corporeal dimensions.

We conclude by emphasizing how the specific use of cognitive neuroscience here proposed can lead to a new take on social cognition. This new take brings about a demonstration on empirical grounds of the constitutive role played in foundational aspects of social cognition by the human body, when conceived of in terms of its motor potentialities; hence its transcendental quality.⁶ Of course this only covers a partial aspect of social cognition. However, we think this approach has the merit of providing an epistemological model, which is also potentially useful for empirical investigation into the more cognitively sophisticated aspects of human social cognition.

2 The cognitive neuroscience of what?

It is a fact and an undisputable truth that there cannot be any mental life without the brain. More controversial is whether the level of description offered by the brain is also sufficient for providing a thorough and biologically-plausible account of social cognition. We think it isn't. We would like to ground this assertion on two arguments: the first deals with the often overlooked intrinsic limitations of the approach adopting the brain level of description, particularly when the brain is considered in isolation and its intimate relation with the body is neglected; the second deals with social cognitive neuroscience's current prevalent explanatory objectives and contents.

The contemporary emphasis divulged by the popular media, namely the supposedly revolutionary heuristic value of cognitive neuroscience, mostly rests upon the results of brain imaging techniques, and in particular on fMRI.

⁶ The "transcendental quality" attributed to the body is intended to mean that the body is considered as the a priori, non-further reducible condition of the possibility of experience.

fMRI is often presented as the ultimate method of investigation of the human mind. It should be pointed out, though, that fMRI studies do not constitute the whole story in cognitive neuroscience. Cognitive neuroscience can indeed carry out its investigation at a more drastic sub-personal level, such as at the level of single neurons (see below), both in macaque monkeys and, although much more rarely, even in humans. These alternative approaches notwithstanding, the main thrust of cognitive neuroscience in studying human brain function is, as we speak, mostly confined to fMRI.⁷

Unfortunately fMRI only indirectly “sees” the workings of the brain, by measuring neurons’ oxygen consumption. Such a measure is also indirect, as it depends on the local difference between oxygenated and deoxygenated hemoglobin—the iron-rich molecule housed by red blood cells, which carries oxygen to all bodily organs and tissues. Oxygenated and deoxygenated hemoglobin have different paramagnetic behaviours in relation to the strong magnetic field that is created by a big coil, inside which the head is placed. The measure of this functional parameter allows scientists to estimate local neural activity in terms of different MRI (magnetic resonance imaging) signals. The indirect quality of this kind of estimation of brain activity, which is based on local hemodynamic brain responses, inevitably introduces distortions and noise. Indeed, when studying any sensorimotor, perceptual, or cognitive function, in order to maximize the so-called “signal-to-noise ratio”, several repetitions of the same task in many individuals are required.

This means that fMRI allows us to indirectly assess the average brain-activation level induced by any given task across a population of no less than twelve to fifteen different individuals. Within each studied individual brain the spatial resolution of fMRI is within the order of few millimetres. This implies that we are able to measure at best the potentially coherent activation pattern of several hundred thousand neighbouring neurons, possibly also differing

among one another in terms of their excitatory or inhibitory role.

Temporal resolution is even worse, since it is in the order of a few seconds. One should consider that action potentials, or “spikes” as neurophysiologists like to call them—the electric code employed by neurons to “communicate” with each other, and ultimately the true essence of neurons’ activity—last less than one millisecond. fMRI cannot match such temporal resolution because it measures the delayed (of about two seconds) and prolonged (for about five seconds) local hemodynamic response providing neurons with all the oxygen their electric activity requires.

As we have previously argued, fMRI is not the only available experimental methodology for studying the brain. Many different techniques are available nowadays (e.g., PET (positron emission tomography), NIRS (near-infrared spectroscopy), Tdcs (transcranial direct-current stimulation) or TMS (transcranial magnetic stimulation)). Particularly, since the revolutionary introduction in 1927 by the Nobel Prize laureate Edgar Adrian ([Adrian & Matthews 1927a](#), [1927b](#)) of the extracellular microelectrode, which allows the recording of action potentials discharged by single neurons, neurophysiology has made enormous progress in revealing the brain’s physiological mechanisms. Such neurophysiological investigation started with the study of the neural circuits that preside over elementary sensorimotor behaviours, like spinal reflexes, finally moving all the way up to the investigation of action and perception, reward and emotions, spatial mapping and navigation, working memory, decision-making, etc., in behaving animals like macaque monkeys. Unfortunately, such a finely grained level of description—both in terms of spatial and temporal resolution, is most of the time precluded in humans.

We posit that the scientific study of intersubjectivity and the human self requires a comparative approach, and that this is the only one capable of connecting the distinctive traits of human nature to their likely phylogenetic precursors. In so doing, and by making use of the single-neuron recording approach, neuro-

⁷ For an intriguing discussion of the historical antecedents of brain imaging techniques and a passionate criticism of the limitations of current use of brain imaging by cognitive neuroscience, see [Legrenzi & Umiltà \(2011\)](#).

physiological mechanisms and the cortico-cortical networks expressing them can be related with several aspects of primates' social cognitive behaviour and thus be thoroughly investigated. Conceptual notions like intersubjectivity and the self should be analyzed in order to better understand their nature, structure, and properties. Such an analysis, which provides us a deflationary notion of the same concepts, intended as the identification of their minimal component and the detailed study of their origin, will be most successful if driven by a meticulous investigation of the underpinning neurophysiological mechanisms—which most of the time are available only from the study of non human primates' brains, hence the necessity of a comparative perspective.

Human brain imaging, because of the intrinsic limitations we briefly outlined above, can only provide correlations between particular brain patterns of activation and particular behaviours or mental states. This implies that the correlation between a particular brain state and a particular phenomenal mental state of a given individual human being⁸ is most informative when the specificity and uniqueness of such a correlation can be firmly established. Unfortunately, this is not always the case with fMRI studies. Very telling is the supposed mindreading specificity of some cortical circuits comprising the ventral portion of the mesial frontal cortex and the TPJ (temporo-parietal junction; e.g., [Leslie 2005](#); [Saxe 2006](#)). Such specificity is not only so far unproven, but is actually confuted by accumulated evidence (for a lengthier discussion of this point and for arguments and experimental evidence against such specificity see [Ammaniti & Gallese 2014](#); [Gallese 2014](#)).

In spite of all these limitations, this neuroimaging approach turned out to be very productive, enabling us to study for the first time in parallel brains, behaviour and cognition, shedding new light not only on human brain structure, but also on its wiring pattern of con-

nectivity and many of its functions. If we put the newly-acquired knowledge on brain function provided by cognitive neuroscience under scrutiny, we can make very interesting discoveries. For example, we discovered that in many areas of investigation brain imaging replicates and validates *at a different scale* what had been previously discovered at the single neuron level in animals like macaques.⁹

The prominent discoveries, among others, of David Hubel, Torsten Wiesel, and Semir Zeki on the functional organization of primates' cortical visual system, like the orientation-, shape-, motion- and colour-selectivity of visual neurons, were made by correlating the discharge activity of single neurons in macaques' visual cortices with different parameters of the visual stimuli macaques were looking at (for a comprehensive review of this literature, see [Zeki 1993](#)). These results later promoted a similar investigation carried out on the human brain by means of fMRI. Remarkably enough, a similar functional architecture was detected in the human visual brain, in spite of the species difference and, most importantly, the different scale at which these investigations were carried out: a few hundreds recorded neurons at best, in the case of macaques' brains, versus hundreds of thousands if not millions of activated neurons detected by a local increase in blood flow in the case of the human brain.

Face-selective neurons, first described in the early nineteen-seventies by Charles Gross and colleagues in macaques' temporal cortex ([Gross et al. 1972](#)), and immediately ridiculed as “grand-mother cells”, offer another very telling example.¹⁰ Face-selective brain circuits appear to be strikingly similar in macaques and humans (see [Ku et al. 2011](#)). Furthermore, even in human brains single neurons were detected that selectively respond to a single face—such

⁸ For sake of concision and focus we do not discuss here the implicated topic of the apparently absent synchronicity between brain states and phenomenal consciousness (remember that fMRI does not provide a good temporal resolution to firmly match brain states and phenomenal consciousness). We simply want to stress the parallel existence of particular experiences and particular brain states.

⁹ The comparative approach is primarily intended here as a comparison between humans and other animals. As a consequence, it also implies a comparison between different experimental methods.

¹⁰ The notion of the grandmother cell was originally introduced by the neuroscientist Jerry Lettvin to refer to neurons with high integrative power, which are able to map concepts or objects. The term has since then mostly been employed with a negative connotation by supporters of a more distributed population-coding of objects, percepts, and memories. For an historical account of the notion of the grandmother cell, see [Gross \(2002\)](#).

as the so-called Jennifer Aniston's selective neurons (see Quiroga et al. 2005). These neurons respond to multiple representations of a particular individual, regardless of the specific visual features of the picture used. Indeed, these neurons respond similarly to different pictures of the same person and even to his or her written or spoken name. The authors of this study claimed that their evidence supported the notion that single neurons within the human medial temporal lobe cortex instantiate the abstract representation of the identity of a single individual.

Such examples seem to suggest that in spite of the big scale magnification implied when confronting single-neuron data from macaques and fMRI results from humans, some important functional features are nevertheless manifest across these different levels of description (i.e., single neurons vs. brain areas). One can study canonical or mirror neurons (see below) by recording the activity of a few hundred spiking neurons from a behaving macaque monkey during object and action observation, respectively. The same results can be replicated by detecting, by means of fMRI, and during object and action observation, the simultaneous activation of hundreds of thousands of human neurons within analogous cortical areas of the human brain.

This remarkable but often neglected fact cannot be the result of a pure coincidence. This evidence should thus invite us to resist and argue against those who downplay the heuristic power of single-neuron recording. Their thesis is that because of a supposedly incommensurable gap between single neurons and the incredible complexity of the human brain, where information would be exclusively mapped at the level of large populations of poorly selective neurons, it doesn't make any sense to study the brain by recording single neurons. The fact however is that in spite of the almost astronomic figures characterizing the human brain (about 100 billion neurons, each of which connects with thousands of other neurons), its complexity does not parallel such astronomic figures, or at least not in such a way as to deny any heuristic value to the single-neuron recording approach. Let's see why.

As argued by Chittka & Niven (2009), brain size may have less of a relationship with behavioural repertoire and cognitive capacity than generally assumed. According to the same authors, larger brains are, in part at least, a consequence of larger neurons that are necessary in large animals due to basic biophysical constraints. Larger brains also contain greater replication of neuronal circuits, adding precision to sensory processes, detail to perception, more parallel processing, enlarged storage capacity, and greater plasticity. These advantages, maintain Chittka & Niven (2009), are unlikely to produce the qualitative shifts in behaviour that are often assumed to accompany increased brain size, or at least not in a one-to-one manner.

The evidence so far briefly reviewed and that we will present in the next sections suggest that some functional properties of the brain exhibit a sort of "fractal quality", such that they can be appreciated at different scales and levels of investigation. For these reasons fMRI cannot be the sole neurocognitive approach to human social cognition,¹¹ but it must be complemented by other approaches compensating for some of its deficiencies: like TMS, EEG (electroencephalography), and the comparative functional study of non-human primates by means of brain imaging and single-neuron recordings.

After having clarified what we take to be the often-neglected limitations of cognitive neuroscience that may hinder its potential heuristic power (namely, that fMRI offers only an indirect estimation of brain activity, inferred by measuring neurons' oxygen consumption, and this inevitably leads to distortions and noise, and that it does not provide good temporal resolution because it measures the delayed and prolonged hemodynamic responses due to neurons oxygenation), let us now move to the second argument against the sufficiency satisfaction condition of the current approach of cognitive neuroscience to the study of human social cognition. This argument concerns the explanatory goals and contents of contemporary main-

¹¹ Maybe no one explicitly claims this. However, it is very common that researchers neglect many (or all) of the complementary techniques that have been here suggested to be necessary and draw inferences about human social cognition and its neural implementation. As an example, see papers by Ian Apperly.

stream cognitive neuroscience. Vast quarters within cognitive neuroscience are still today strongly influenced by classical cognitivism, on one side, and by evolutionary psychology on the other. Classical cognitive science is the bearer of a solipsistic vision of the mind, according to which focusing on the mind of the single individual is all that is required in order to define what a mind is and how it works. The image of the mind that classical cognitive science gives us is that of a functional system whose processes are described in terms of manipulations of informational symbols in accordance with a series of formal syntactic rules.

According to evolutionary psychology, by contrast, the human mind is a set of cognitive modules, each of which has been selected during evolution for its adaptive value. Major figures of this current, such as John Tooby and Leda Cosmides, have gone as far as maintaining that the brain is a physical system that works like a computer (Cosmides & Tooby 1997). According to Steven Pinker (1994, 1997), our cognitive life can be referred to in terms of the function of a series of modules like the linguistic module, the module for the Theory of the Mind, etc.

Based on this theoretical framework, in the last twenty years cognitive neuroscience when investigating human social cognition has mainly tried to locate—as mentioned above—the cognitive modules in the human brain. Such an approach suffers from ontological reductionism, because it reifies human subjectivity and intersubjectivity within a mass of neurons variously distributed in the brain. This ontological reductionism chooses as a level of description the activation of segregated cerebral areas or, at best, the activation of circuits that connect different areas and regions of the brain. However, if brain imaging is not backed up by a detailed phenomenological analysis of the perceptual, motor, and cognitive processes that it aims to study and—even more importantly—if the results are not interpreted, as previously argued, on the basis of the study of the activity of single neurons in animal models, and the study of clinical patients, then cognitive neuroscience, when exclusively consisting in brain imaging, loses much of its heuristic power. Without the

demonstration of the specific correlation between brain states and mind states and the explanation of such correlation, much of the contemporary brain imaging approach to social cognition looks like a sort of high-tech version of phrenology.

For this reason a “phenomenologization” of cognitive neuroscience is desirable, as Gallese has proposed before (see Gallese 2007, 2009, 2011, 2014). In Gallese’s view, to “phenomenologize” cognitive neuroscience means to start neuroscientific research from the analysis of subjective experience and of the role that the living body plays in the constitution of our experience of material objects and of other living individuals. In this way, the empirical study of the genetic aspects of subjectivity and intersubjectivity can be pursued on new bases—if compared to those thus far adopted by classical cognitivism. Francisco Varela a few years ago realized a similar possibility and set out on a pathway of analysis in this direction (Varela & Shear 1999).

Times change, however. We are insisting on nothing less than a change of paradigm. A new neuroscientific approach to the study of the human mind is gaining momentum. It capitalizes upon the study of the bodily dimension of knowledge: the so-called “embodied cognition” approach. In the next section we introduce mirror neurons and embodied simulation. Our purpose is to show that, starting from a sub-personal neuroscientific description of the pragmatic relationship with the world, a pathway can be traced to define the forms of subjectivity and intersubjectivity that distinguish human nature, rooted in the bodily interacting nature of human beings.

3 Mirroring mechanisms and embodied simulation

The discovery in the early 1990s of mirror neurons in the brain of macaques (Gallese et al. 1996; Rizzolatti et al. 1996), and the subsequent discovery of mirror mechanisms in the human brain (see Gallese et al. 2004; Rizzolatti & Sinigaglia 2010) suggest that there exists a direct modality of access to the meaning of other

people's behaviours—a modality that can be set aside from the explicit attribution of propositional attitudes. Mirror neurons are motor neurons, originally discovered in macaques' ventral premotor cortex area F5, later on also found in the reciprocally-connected posterior parietal areas AIP and PFG. Mirror neurons not only respond to the execution of movements and actions, but also respond to the perception of actions executed by others. Mirror neurons map the action of others on the observers' motor representation of the same action. Further research also demonstrated in the human brain the existence of a mechanism directly mapping action perception and execution, defined as the Mirror Mechanism (MM; for a recent review, see [Ammaniti & Gallese 2014](#); [Gallese 2014](#)). In addition, in humans the motor brain is multimodal. Thus, it doesn't matter whether we see or hear the noise made by someone cracking peanuts, or locking a door. Different—visual and auditory—sensory accounts of the same motor behaviour activate the very motor neurons that normally enable it. The brain circuits showing evidence of the MM, connecting frontal and posterior parietal multimodal motor neurons, most likely analogous to macaques' mirror neurons, map a given motor content like “reach out” or “grasp” not only during their performance, but also when perceiving the same motor behaviour performed by someone else, when imitating it, or when imagining performing it while remaining perfectly still. The relational character of behaviour as mapped by the cortical motor system enables the appreciation of purpose without relying on explicit propositional inference.

Altogether, these findings led to the formulation of the “Motor Cognition” hypothesis as a crucial element in the emergence of social cognition ([Gallese 2009](#)). According to this hypothesis, cognitive abilities like the hierarchical representation of action with respect to a distal goal, the detection of motor goals in others' behaviour, and action anticipation are possible because of the peculiar functional architecture of the motor system, organized in terms of goal-directed motor acts. Traditionally, the relation between actions and their outcomes is assumed

to be largely independent of the motor processes and representations underpinning action execution. Such processes and representations allegedly concern elementary motor features such as joint displacements or muscle contractions only. However, solid empirical evidence challenged this view. Motor processes may involve motor representations of action goals (e.g., to grasp, to place, etc.), and not only kinematic or dynamic components of actions. This suggests that beliefs, desires, and intentions are neither primitive nor the only bearers of intentionality in action. We do not necessarily need to metarepresent in propositional format the intentions of others to understand them. Motor outcomes and motor intentions are part of the “vocabulary” that is spoken by the motor system. On occasion we do not explicitly ascribe intentions to others; we simply detect them. Indeed, we posit that motor representation is enough to ground the directedness of an action to its outcome ([Gallese 2000](#), [2003b](#); [Butterfill & Sinigaglia 2014](#); compare also [Gallagher's 2005](#) notion of direct perception).

One of the consequences of the discovery of mirror neurons was the possibility of deriving subjectivity from intersubjectivity at the subpersonal level of description. The sense of self is precociously developed, beginning from a self that is first of all physical and bodily, and which is constituted precisely by the possibility of interacting and acting with the other. Embodied simulation can provide the neurobiological basis for early forms of intersubjectivity, from which the sense of the self is built. The discovery of mirror neurons and the simulation mechanism would therefore seem to further stress that being a self also implies being with the other. The model of intersubjectivity suggested by mirror mechanisms and embodied simulation correlatively sheds new light on the subjective dimension of existence. Let us see first what type of intersubjectivity mirror neurons seem to suggest.

The discovery of mirror neurons gives us a new empirically-grounded notion of intersubjectivity connoted first and foremost as intercorporeality—the mutual resonance of intentionally meaningful sensorimotor behaviours. The ability

to understand others as intentional agents does not *exclusively* depend on propositional competence, but it is highly dependent on the relational nature of action. According to this hypothesis, it is possible to directly understand the meaning of other people's basic actions thanks to a motor equivalence between what others do and what the observer *can* do. Intercorporeality thus becomes the main source of knowledge that we have of others. The motor simulation instantiated by neurons endowed with "mirror properties" is probably the neural correlate of this human faculty, describable in functional terms such as "embodied simulation" (Gallese 2003a, 2005, 2011; Gallese & Sinigaglia 2011b).

Action constitutes only one dimension of the rich baggage of experiences involved in interpersonal relations. Every interpersonal relation implies the sharing of a multiplicity of states like, for instance, the experience of emotions and sensations. Today we know that the very nervous structures involved in the subjective experience of emotions and sensations are also active when such emotions and sensations are recognized in others. A multiplicity of "mirroring" mechanisms is present in our brain. It was proposed that these mechanisms, thanks to the "intentional attunement" they generate (Gallese 2006), allow us to recognize others as our fellows, likely making intersubjective communication and mutual implicit understanding possible. The functional architecture of embodied simulation seems to constitute a basic characteristic of our brain, making possible our rich and diversified intersubjective experiences, and lying at the basis of our capacity to empathize with others.

4 Body and self

After having delineated a deflationary neurobiologically-grounded account of basic aspects of intersubjectivity, namely an account focused on the minimal core mechanisms of intersubjectivity, let us now address the relationship between body and self. A minimal manifestation of the sense of self can already be identified in our first bodily experiences, and this highlights the potential contribution of bodily experiences to

its constitution. Some aspects of the minimal self proposed by contemporary philosophical and empirical research are the notion of first-person perspective, the "mineness" of the phenomenal field (*Meinigkeit*), embodiment of point of view, and issues of agency and body ownership (Cermolacce et al. 2007).¹² On the philosophical side, phenomenology emphasizes the necessity of embodiment of the self for all the above-cited aspects of self experience. As argued by Cermolacce et al. (2007, p. 704, footnote 3), in phenomenology

the field of experience is not yet considered to be subjective because this predicate already implies that there is a subject. For phenomenology, the very idea of the subject articulates itself in experience. In this sense, the manifestation and appearing of experience are the conditions for the experience of the subject in question.

This philosophical standpoint has important implications for the empirical investigation of the neural correlates of the self.¹³ Rather than empirically addressing the self by starting with a search for the neural correlates of a pre-defined, explicit, and reflective self-consciousness, we believe it to be more productive to investigate what set of constitutive conditions allows an implicit and pre-reflective sense of self to emerge, and how this is effected. The interesting questions to be first answered are: "What enables the basic experience of ourselves as bodily selves? What enables us to implicitly distinguish ourselves, as bodily selves, from other human bodily selves?" In the following we review and discuss recent empirical evidence providing preliminary answers to these questions.

¹² Again, for sake of concision, we do not deal here with the relationship between the notion of a core, minimal self as a bodily self and agency and body ownership. On this topic, see Gallese & Sinigaglia (2010, 2011a). Moreover, it is worth noting that the arguments proposed in this section in relation to the notion of a minimal bodily self could be applied to other non-human animals. The possibility that high-level self-awareness can emerge from a primitive and non-conceptual form of self-awareness, and that it is possible that we share this basic level of the sense of self with other non-human animals, has already been discussed. For a discussion of this and other related topics see Bermúdez (2003).

¹³ See Vokey et al. (2003) and Vokey et al. (2004) for an investigation of the neural correlates of the first-person perspective.

The relationship between the minimal sense of self and the cortical motor system was recently revealed. The motor experience of one's own body, even at a covert level, allows an implicit and pre-reflective bodily self-knowledge to emerge, leading to a self/other distinction. Indeed it was recently shown that in a task in which differently rotated static pictures of right and left human hands were presented, participants who had to determine whether each observed hand was the right or the left one produced faster responses when observing the pictures of their dominant hand with respect to others' hands (Ferri et al. 2011). However, when participants were asked to explicitly discriminate between their hands and the hands of others, the self-advantage disappeared. Implicit and explicit recognition of the bodily self dissociated: only implicit recognition of the bodily self, mapped in motor terms, facilitated implicit bodily self-processing.

A subsequent fMRI study by Ferri et al. (2012), using a similar hand mental rotation task, demonstrated that a bilateral cortical network formed by the supplementary and pre-supplementary motor areas, the anterior insula, and the occipital cortex was activated during processing of participants' own hands. Furthermore, the contralateral ventral premotor cortex was uniquely and specifically activated during mental rotation of the participants' own dominant hands. The ventral premotor cortex might represent one of the essential anatomo-functional bases for the motor aspect of bodily selfhood, also in light of its role in integrating self-related multisensory information. This hypothesis is corroborated by clinical and functional evidence showing its systematic involvement with body awareness (Ehrsson et al. 2004; Berti et al. 2005; Arzy et al. 2006). This evidence demonstrates a tight relationship between the bodily self-related multimodal integration carried out by the cortical motor areas, specifying the motor potentialities of one's body and guiding its motor behaviour, and the implicit awareness one entertains of one's body *as* one's own body and of one's behaviour *as* one's own behaviour. Because the ventral premotor cortex is anatomically connected to visual and somato-

sensory areas in the posterior parietal cortex and to frontal motor areas we hypothesize that premotor cortex activity, by underpinning the detection of congruent multisensory signals from one's own body, could be at the origin of the experience of owning one's own body parts.

This minimal notion of the self, namely the bodily self as power-for-action (see Gallese & Sinigaglia 2010, 2011a), tacitly presupposes ownership of an action-capable agentive entity; hence it primarily rests upon the functionality of the motor system. As we just saw, empirical evidence supports the neural realization of this implicit aspect of selfhood in the brain's motor cortex. Since the minimal bodily self rests neurally on the motor system, it logically follows that characteristics of the latter are defining for the former. This implies that one could attribute to the minimal bodily self known features of the motor system, including its capacities and limitations. The motor aspects of the bodily self provide the means to integrate self-related multimodal sensory information about the body and the world with which it interacts. This is also important from a theoretical point of view, because it opens the possibility of linking the openness of the self to the world to the motor potentialities its bodily nature entails.

One could then posit that the minimal bodily self when conceived in terms of its motor potentialities has a dual function. On the one hand, it constitutes important aspects of the basic sense of self. On the other, it shapes our perception and pre-reflective conception of others as other selves incarnated in a motorly-capable physical body with capacities and experiences similar to ours. Through mirror mechanisms and embodied simulation, others appear to us as second selves, or second persons. We believe that this perspective provides a more vivid experience of intersubjectivity, relative to the detached, propositional deliberation on the experiences of others available in standard mind reading of others.

5 Body and language: Reflexiveness

According to the perspective so far delineated, body, actions, and feelings play a direct role in

our knowledge of others. The question remains open as to whether our propositional representations are totally separate from this bodily dimension. Our hypothesis is that they are not. But it remains a fact that linguistic and bodily cognition afford us diversified modalities of epistemic access to the world, even though often such modalities contaminate one another and are inevitably interwoven.

The mind, from the perspective delineated here, is therefore an embodied mind, though it would be more correct to speak of a corporeal mind. The concept of embodiment can induce one to think that a mind pre-existing the body can subsequently live in it, and use it. The truth is that mind and body are two levels of description of the same reality, which manifests different properties according to the chosen level of description and the language employed to describe it. A thought is neither a muscle nor a neuron. But its contents, the contents of our mental representations, are inconceivable without our corporeity. Likewise it is difficult to imagine how the representational format of a propositional type can have developed without our corporeality. Language somehow allows us, as we will see, to transcend our corporeity; nevertheless, we posit that the bond with the body is always present.

A few years ago, [Gallese \(2000\)](#) proposed that we look at the evolution of human language as an exaptation¹⁴ of functional sensorimotor processes, which put them into the service of human linguistic competence. The hypothesis of exaptation was then developed in subsequent papers and later elaborated in terms of “neural exploitation” ([Gallese & Lakoff 2005](#); [Gallese 2008](#)), or “neural reuse” ([Gallese 2014](#)). “Neural exploitation” consists in the reuse of neural resources, originally evolved to guide our interactions with the world, to serve the more recently evolved linguistic competence. This notion of reuse implies a functional uncoupling of the sensorimotor system from muscular output, to guide the generative-syntactic aspects of language by functionally connecting it to the pre-

frontal and, more generally, non-sensorimotor circuits. According to this view, intentionality, the aboutness of our representations, is—in the first place—an exapted property of the action models instantiated by the cortical motor system ([Gallese 2000](#), p. 34). The sensorimotor system, when uncoupled from muscular output, makes available to us a model, or paradigm, of our motor knowledge. As such, not only it houses causative properties but also content properties. And this relation to a content, or aboutness, is a primitive expression of intentionality, then exploited by other forms of representations. This perspective on reuse is acquiring more and more supporters (see [Dehaene 2005](#); [Anderson 2010](#)).¹⁵

Compelling evidence shows that humans, when processing language, activate the motor system both at the phono-articulatory and at the semantic level. When listening to spoken words or looking at someone speaking to us, our motor system simulates the phono-articulatory gestures employed to produce those very same words. Furthermore, processing action-related linguistic expressions activates regions of the motor system congruent in somatotopic fashion with the processed semantic content. Reading or listening to a sentence describing a hand action activates the motor representation of the same action (for a review, see [Gallese 2008](#); [Glenberg & Gallese 2012](#)). Interestingly, somatotopic motor activation has also been observed during the comprehension of abstract and figurative use of language such as metaphors and idioms (e.g., [Guan et al. 2013](#); [Boulenger 2012](#); see also [Gallese & Lakoff 2005](#) on the bodily foundation of concepts). However, it is important to note that embodied simulation is not always involved in language comprehension, and that there is no contradiction in saying this. There are cases in which language, at least at the content level, is not tied to any form of bodily knowledge. In such cases (e.g., when we talk about the notions such as moral judgement or intelligence) no embodied simulation is likely to be at play.

Nevertheless, the problems that language raises for the embodied perspective on human

¹⁴ Exaptation refers to the shift in the course of evolution of a given trait or mechanism, which is later on reused to serve new purposes and functions (see [Gould & Lewontin 1979](#)).

¹⁵ For a discussion of different views on the notion of reuse, see [Gallese \(2014\)](#).

social cognition are still enormous. As clearly underlined, among others, by the Italian philosopher [Paolo Virno](#) (2003, 2011), the common linguistic space shared by a community of speakers proves to be incommensurably different from the pre-linguistic one. The linguistic dimension is based on a distinction between linguistic utterances and facts about the world, be they referable to physical or psychological events. We can say that “today the sun is shining” and be understood, even if outside the window snow is falling. Or we can maintain that “all Italians want to pay taxes”, again being understood and simultaneously contradicted by the factual truth of the enormous tax evasion in our country.

According to Virno, the gap between meaning and denotation (what he calls the neutrality of meaning, namely the fact that the meaning of a word such as, for example, “man”, can be understood apart from any reference to an instance of man) is referable to linguistic reflexivity, i.e., to the fact that language refers to itself and that with words we can talk about other words. It seems to us that the reflexivity of language is a product of the symbolic nature of linguistic representations. The symbolic nature of such representations is what allows language to break away from the “here and now”; it is what allows the neutrality of meaning.

In order for a sign to be symbolic it necessarily has to be reflexive. What makes a sign a symbol is its being part of a system in which each term is correlatively defined in relation to the other terms within the system and in relation to the renegotiation of this relationship constantly taking place within the system itself. It is the use of a symbol within a given context that each time redefines relationships inside the language system.

Thus, symbolic relationships are by definition characterized by reflexivity. Symbols are defined through other symbols. This level of reflexivity is pre-theoretical; it emerges in the linguistic activity of each speaker and leads to a form of linguistic awareness of a practical character. This practical linguistic awareness has been called the *epilinguistic* quality and thus it

has been distinguished from the theoretical quality that is expressed in the metalanguage of linguistics ([Culioli 1968](#); [Lo Piparo 2003](#)). The concept of epilinguistic quality refers to the natural tendency of speakers to reflect on their own language—a tendency made possible by the distinctive quality of language being able to speak of itself.

The uniqueness of human language is also maintained by classical cognitivism and by cognitive linguistics, but for very different reasons. The otherness of human language in comparison with other systems of communication known in the animal world derives from its linguistic recursive quality. In an often-quoted article written some years ago ([Hauser et al. 2002](#)) defined the faculty of language in a narrow sense (FLN) as being expressed by recursivity. Nevertheless, this perspective, in addition to suffering from the usual cognitivist solipsism, is exposed to comparative verification in the animal world. If the FLN marks human linguistic uniqueness in terms of syntactic recursivity, the latter must be entirely absent in the extra-human animal kingdom.

Actually, the facts tell us exactly the opposite. Recent studies ([Gentner et al. 2006](#); [Abe & Watanabe 2011](#); see also [Margoliash & Nusbaum 2009](#); [Bloomfield et al. 2011](#)) have shown that singing species of birds like starlings or finches demonstrate, both in the production and in the reception of conspecifics’ vocalizations, the ability to produce and to extract recursive syntactic characteristics. The study by Abe and Watanabe also shows that the development of this competence is dependent on social encounters with the vocalizations of other conspecific individuals. Finally, these authors have shown that lesion of the lateral magnocellular nucleus of the anterior nidopallium, a motor structure comparable to the basal ganglia of primates, involved both in the production and the perception of song, prevents finches from discriminating the syntactic-recursive characteristics of the song they hear.

These results show that the best strategy for studying some of the most relevant aspects of human social cognition, even demonstrating the bases of their uniqueness, consists in a pre-

liminary recognition of the mechanisms and faculties that we share with the rest of the animal world. As maintained in the past (Gallese 2003b, 2008), the difference between human and nonhuman nature could originally have been of a quantitative rather than a qualitative nature.¹⁶

6 Body and language: Facts and challenges

One of the key challenges for the embodied approach to human social cognition consists in trying to understand whether and how our bodily nature determines some of our linguistic activities, such as denying, asking, or doubting, that seem to be exclusively human. Are linguistic activities as those ones anchored to bodily mechanisms? The question is open and empirical research must address this challenge in the coming years. In the meantime, at least at a purely speculative level, let us try to delineate a possible point of contact between the anthropogenic power of language and embodied simulation.¹⁷

There is indeed a way to connect the common pre-linguistic sphere to the linguistic one (Gallese 2003b, 2007, 2008; Gallese & Lakoff 2005; Glenberg & Gallese 2012). This consists in showing that language, when it refers to the body in action, brings into play the neural resources normally used to move that very same body. Seeing someone performing an action, like grabbing an object, and listening to or reading the linguistic description of that action lead to a similar motor simulation that activates some of the same regions of our cortical motor system, including those with mirror properties, normally activated when we actually perform that action.

These data on the role of simulation in understanding language (see Pulvermüller 2013 for a review of this topic) broadly confirm a thesis already discussed in the history of philosophy (for instance, by Epicurus, Campanella, Vico,

see Usener 1887 and Firpo 1940 or Condillaco 2001). The thesis in question claims for the bodily, sensory, and motor dimensions a constitutive role in language production and understanding. However, it seems that the relationship between language and body does not move in a single direction. The fact is that language is without doubts constitutive of human nature and, as such, it seems to offer us wholly human modalities of experiencing our corporeity.

In this sense, neuroscientific data on the role of simulation during understanding of language also lend themselves to a mirror and complementary reading with respect to that previously proposed. On the one hand it is plausible that embodied simulation might play a crucial role in understanding language. Indeed, if one reversibly interferes with this process, for instance by means of TMS stimulation, understanding of language is jeopardized. On the other hand, language allows us—and in this we are unique among all living species—to fix and relive specific aspects of our bodily experience. Through language we can crystallize and relive fragments of experiences that are not topical, that is to say are not *my* experiences *now*, but become a paradigm, a model, for understanding ourselves and others.

In the following section we discuss the role of embodied simulation seen as a *paradigm* or model in the light of the Aristotelian notion of *paradeigma*.¹⁸ For the time being it suffices to stress that the possibility of hypostatizing and then segments of our experiences independently of our immediate physical context, or independently of specific physical stimuli, is a possibility that only the possession of language allows us to experience. The faculty of language is therefore, on one side, rooted in corporeality but, in turn, changes and moulds our way of living bodily experiences.

7 Body and language: Embodied simulation as a paradigm?

The relation between body and language was to a great extent underestimated in the last cen-

¹⁶ According to this perspective, linguistic syntax could originate and be modelled upon syntactic motor competence, the latter being adapted and put at the service of the new linguistic competence (see Gallese 2007, 2008).

¹⁷ With the expression “anthropogenic power of language” we mean that the human nature, as we know it, depends on language.

¹⁸ For an earlier formulation of this hypothesis, see Gallese (2013).

tury, thanks, above all, to Chomsky's major influence. In 1966 Chomsky published a book significantly entitled *Cartesian Linguistics*. Descartes is the originator of the idea that language has little to do with the body.¹⁹ The Cartesian thesis on the relationship between language and body implies, on one side, that the body is not a substratum and material of language and, on the other, that language is exclusively a tool that expresses a thought formed independently of language itself. According to Descartes (1642) and the Cartesian tradition in which Chomsky stands, language is a tool through which we manifest an autonomous thought that precedes language—a thought structured by logic but certainly not by language, whose role is circumscribed and downsized to that of being a mere label of thoughts (cf. Hinzen & Sheehan 2013 for a critical discussion of the issue).

The theses informing the Cartesian idea of language are challengeable nowadays. Language makes meaning general, releasing it from the context, that is, from the dimensions of who, what, how, where, and when. Language, in other words, provides us with a unique modality of reference to the world, allowing us at the same time to transcend contingent determinations and to define them at a different level, thanks to the use of concepts like subject, object, time, space, universal, etc. It is perhaps not trivial to notice that such concepts correspond to precise grammatical structures and that, most likely, the use of a grammatically-structured language contributed, by co-evolutionary dynamics, to the structuring of rational thought characterized by such features (Hinzen & Sheehan 2013).

Hence, thanks to language we can speak of humankind without referring in particular to any of the single individuals sharing the property of belonging to the human species. We can speak of a subject aside from the individual embodiments of this attribute, etc. Language, as stressed by Virno, furnishes us with general meanings, that is, meanings valid

for everybody but, at the same time, meanings that do not necessarily denote a particular instantiation.

Interestingly enough, according to Giorgio Agamben (2008) what holds “for everybody and nobody” is referable to the Greek notion of *paradeigma*, originally explored by Aristotle. The *paradeigma* is a typical form of rhetorical reasoning that moves between individual and individual according to a form of bipolar analogical knowledge. Agamben (2008, pp. 23-24), radicalizing Aristotle's theses, maintains that the paradigm can only be conceived of by abandoning the dichotomy between individual and universal: the rule does not exist before the single cases to which it is applied. The rule is nothing but its own exhibition in the single cases themselves, which thus it renders intelligible.

By applying the notion of paradigm to the grammatical “rules” of language, Agamben touches upon a central point: the so-called linguistic rule derives from the suspension of the concrete denotative application:

[t]hat is to say, in order to be able to serve as an example, the syntagm must be suspended from its normal function, and, nevertheless, it is precisely through this non-operation and this suspension that it can show how the syntagm works, can allow the formulation of the rule. (Agamben 2008, p. 26)

To better explain the notions of rule and suspension of a denotative application, Agamben refers to Latin declensions. When we want to learn the first declension we inflect a noun, for example “*rosa*”, “*rosae*”, etc... In so doing, we are suspending the usual denotative application of this noun and, by means of this suspension, we are showing how the declension works. According to Agamben, “[...] in the paradigm, intelligibility does not precede the phenomenon, but is, so to speak, ‘alongside’ it (*parà*)” (2008, p. 29). In other words “[...] in the paradigm there is not an origin or an *arché*: every phenomenon is the origin, every image is archaic” (Agamben 2008, p. 33).

¹⁹ It is worth noting that Descartes also defended the related thesis that animals don't have soul exactly because they do not have language. Cf. Descartes (1637).

On Agamben's reading, the Aristotelian *paradeigma* is a good model for describing the creation of linguistic rules. Starting from Agamben's intuition and seeking to move one step further, the hypothesis that we want to explore here is that the notion of *paradeigma* is a good model not only for the creation of linguistic rules but also for the definition of the embodied simulation mechanism. In this connection, simulation allows us, at a sensorimotor level, to hypostatize and reuse what holds "for everybody and nobody".

To understand to what extent the analogy between embodied simulation and *paradeigma* works it is necessary to go back to Aristotle. What is meant by *paradeigma* in Aristotelian thought and in what context does Aristotle make use of this notion?

The *paradeigma*, as already anticipated, is a typical is a typical argument form used to persuade and devoted to the discussion of "things that can be otherwise." Aristotle discusses this argument form, which does not have any demonstrative aim, both in *Prior Analytics* and in *Rhetoric*. Argumentation based on the *paradeigma*, for example, consists in the presentation by the orator of an exemplary case, based on a historical fact or a figment of the imagination, as in the case of fables. It is the juxtaposition of the present situation and an exemplary one that guides, or should guide, the actions of the person to whom the argumentation is addressed. Thus the *paradeigma*, among rhetorical argumentations, is that which goes from the particular to the particular, from an exemplary case to the present situation. Argumentation based on the *paradeigma* does not make a claim for universality. The orator is not bound to offer an exhaustive number of cases justifying a universally valid conclusion. One case is sufficient, provided that it is particularly suitable, and precisely exemplary, in relation to the context in which the argumentative discourse takes place.

For these reasons, though resembling inductive reasoning (*epagoghé*), which proceeds from the particular to the universal, and indeed considered by Aristotle himself as the transposition of inductive reasoning to the rhetorical

sphere, the *paradeigma* conquers its own autonomous space. To confirm this, one need only to think that in *Prior Analytics* Aristotle (Ross 1978) devotes two separate chapters to paradigm and induction: respectively XXIV and XXI of Book II.

On the one side the *paradeigma*, which proceeds "from the part to the part" (*Prior Analytics* 69a, 15), are peculiar aspects distinguishing it from the *epagoghé*; on the other, it is by all means a form of induction, as Aristotle expressly affirms at the start of Chapter XX of Book II of *Rhetoric*. According to Piazza (2008, p. 117) there are at least two reasons why the *paradeigma* can still be considered a form of *epagoghé*, despite the peculiarities that characterize it. Both these reasons seem interesting, not only for the definition of paradigm that, starting from Aristotle, Agamben discusses in relation to linguistic praxis, but also and above all in the framework of a reflection on the mirror mechanisms enacted in embodied simulation.

Following Piazza's (2008, p. 117) reading of Aristotle, the first of the characteristics of inductive reasoning also found in the *paradeigma* consists in always proceeding from what is "best known and first for us" (Aristotle, *Analytica posteriora* II.19), or from what is for us most immediate and most easily accessible, because being part of our baggage of experiences and knowledge. The second characteristic is, instead, identifying similarities between particular cases.

At another level of analysis, both these features also characterize embodied simulation. One condition for the simulation mechanism's being enacted is sharing a baggage of (motor) experiences and knowledge. Embodied simulation is enacted starting from what for us is "first," i.e., what for us is known and easily accessible in terms of motor potentialities and experiences. Sharing a repertoire of practices, experiences, and sensations is therefore an essential condition, since only by starting from what is well known to us it is possible to identify analogies between our actions and others'. We understand the other starting from our own bodily experience, which is what is "best known and first for us", again using Aristotle's words. On

the basis of this knowledge we identify similar elements in our experiences as well as in those of others.

Embodied simulation, when manifested in the phenomenon of action, emotion, or sensation *mirroring* always involves an original I-thou relationship in which the “thou” is the term with respect to which the self is constituted. On the other hand, the “self” is the basis on which immediate and implicit understanding of the “thou” is possible.

The analogy with the cognitive mechanism subtended by paradigmatic reasoning appears evident. Indeed, in the case of Aristotle’s *paradeigma*, an example, a particular case, is understood because it is close to our feeling, our experiences, and our baggage of knowledge. And nevertheless the process does not stop here. This form of understanding of a particular that is not me will lead me to new conclusions and to a deeper understanding of myself, of *my* particular case, and of *my* situation. Our experiences are therefore the measure from which we understand others and their experiences (i.e., our previous actions, emotions, and so forth). And others’ experiences (i.e., their actions, emotions and so forth) are for us a condition for deeper understanding of ourselves. Thus, the embodied simulation underpinning my present experience is also a *paradeigma* from which I can understand what I observe in others and draw inferences from it for others and for myself.

The embodied simulation mechanism, thus defined, is constitutive of the process of construction of meaning. In this connection, embodied simulation enacted while understanding language is not my present experience but the *paradeigma* in relation to which some of our linguistic expressions acquire a meaning that is rooted in the body. When we read or listen to the description of an action, the process of simulation that takes place in us is not the enactment of the same action; we would be echopractic if we were unable to avoid imitating and reproducing all the actions that we see or whose description we listen to or read. According to our hypothesis, embodied simulation rather makes available to us an exemplary case, a

model, in relation to which understanding of language is also realized. If therefore it is true that the symbolic dimension opens up some possibilities for us and creates worlds for us that only linguistic creatures can enter, it is also true that language strongly exploits mechanisms rooted in our corporeality. Enactment of the simulation process in understanding language seems to suggest that the symbolic dimension and the bodily dimension cohabit in linguistic praxis.

Nevertheless, the nature of this relationship is still not entirely clear, nor is the confine between the bodily dimension and the typically or exclusively symbolic dimension. Can it be hypothesized that corporeal knowledge also plays a role in understanding logical operators such as, for instance, negation or disjunction, or that it plays a role in understanding the interrogative form? The whole symbolic nature of these linguistic structures appears in some respects beyond question. Research on these issues is now open (Kaup et al. 2006, 2007; Tettamanti et al. 2005; Christensen 2009; Tomasino et al. 2010; Liuzza et al. 2011; Kumar et al. 2013) and today many wonder about the possibility of identifying mechanisms that can anchor such structures to our bodily experience. We take this to be the real challenge for the embodied cognition approach to the role played by language in human social cognition.

Let us once more return to the Aristotelian notion of *paradeigma* and appraise other possible hints for substantiating the analogy with the embodied simulation mechanism. The understanding that the rhetor calls for through reasoning based on the *paradeigma* should lead the citizen to choose what is best for him in various circumstances. The goal of such reasoning is to determine understanding of a present situation, by analogy with a historical example or a fable, and, on the basis of this more informed knowledge, to guide the human being’s choices. In other words, the rhetorical example or *paradeigma* is knowledge whose main goal is practical and not theoretical.

A practical aim also characterizes embodied simulation. Embodied simulation is always aimed at “navigating” in the world and, there-

fore, eventually at acting. It was hypothesized that embodied simulation allows us a direct, experiential way of understanding other people's actions and experiences and, on the basis of this understanding, it allows us to regulate our actions and our experiences. These goals are always practical. In some respects, the process of embodied simulation that is enacted, for instance, when reading a novel (see Wojciehowski & Gallese 2011), also has a practical aim. Literature recreates a world of emotions and experiences: the emotions and the experiences of the literary characters inhabit the fictional world of the novel. The simulation mechanism helps us to "navigate" that world, even if it is a fictitious world; it allows us to understand and, in part, to relive the emotions of the protagonists and their vicissitudes. The aim in this case is practical insofar as the simulation mechanism allows us to approach the fictitious other with a second-person epistemic perspective (Gallese 2014).²⁰

Embodied simulation makes implicit knowledge about others immediately available, with the aim of regulating our interactions with them. Our understanding of the literary other is almost always second-person, based on the possibility of perceiving analogies between our own experiences and others' and made possible through a hypostatization of our experiences that is achieved through the simulation mechanism (Wojciehowski & Gallese 2011).

In the end, what is embodied simulation if not a suspension of the application of a process? Let us think of when mirror neurons are activated in observing actions performed by others; or of when canonical neurons are activated while we are looking at the keyboard of a computer thinking about what we want to write; or when cortical motor neurons are activated when we imagine ourselves writing on that keyboard. These responses on the part of motor neurons manifest the activation of implicit knowledge, that is, bodily motor knowledge expressing the motor potentialities of the bodily self mapped by the motor system in terms of their motor outcomes.

Reuse of motor knowledge, in the absence of the movement that realizes it, as exemplified by embodied simulation, is an example of "paradigmatic knowledge." Thus, embodied simulation is a case of implicit paradigmatic knowledge. According to our hypothesis, embodied simulation allows us to naturalize the notion of paradigm, anchoring it at a level of sub-personal description, whose neural correlates we can study.

Our openness to the world is constituted and made possible by a motor system predisposing and allowing us to adapt our daily and contingent pragmatic relationships with the world against the background of a prefigured but highly flexible plan of motor intentionality. Such a plan, intended as the sum of our motor potentialities, provides its coordination to any single contingent modality of relation with the world, that is, to any single action we perform, in which it continues to actualize itself. This aspect seems important to us because it shows how specific aspects of human social cognition are made possible and scaffolded upon processes not necessarily specific to humans, like embodied simulation.

8 Body and soul

It is improper, or at least it seems so to us, to say that the soul, the spirit, or intelligence are embodied. If it were so, we would thus return to a dualistic conception of human nature. Such dualism is always present in the tradition of Western thought, though it is often disguised in forms that, rightly or wrongly, are deemed politically more correct. Today nearly all cognitive scientists declare themselves to be monists and physicalists. Nevertheless, the conception still dominant today about the cognitive structure of humans and their functions draws a clear-cut and apparently unsolvable division between linguistic-cognitive processes and sensory-motor processes. It matters little that everybody admits that both are in some way referable to the biological physicality of the brain. The brain, according to classical cognitivism, is divorced from the body and conceived of as a box of algorithmic wonders.

Classical cognitivism sees the body as an appendix of little or no interest for decoding the supposed algorithms reportedly presiding over

²⁰ A second-person perspective is adopted in social contexts when, implicitly or explicitly, we re-use our own experiences to understand others. On the notion of second-person perspective see Pauen (2012).

our cognitive life. Such is a cognitive life with very little vitality, totally divorced not only from effectual reality, as we try to show here, but also from our daily phenomenal experience. It is not by chance that the language usually used to describe cognitive processes is borrowed from artificial intelligence: algorithms, information processing, etc.

Humans, however, cannot be assimilated to information-processing entities. Even less acceptable is the thesis that the concept of meaning is wholly assimilable to the concept of information. Classical cognitivism has maintained for decades that intelligence depends on the algorithms that substantiate it and not on the material substrata on which the algorithms themselves are believed to be implemented. This is the so-called principle of the multiple realizability of cognitive processes. Embodied Simulation and its relation to language and cognition casts severe doubts on this principle and adds arguments in support of the thesis that human cognition is tightly and necessarily dependent on the kind of body we have. As such, the mechanism of Embodied Simulation and the role it plays in human cognition provide further arguments in support of the idea that the principle of multiple realizability is false. We are who we are because we evolved by adapting to a physical world that obeys a series of physical laws, such as that of gravity.

As the art historian [Heinrich Wölfflin](#) wrote in his 1886 *Prolegomena zu einer Psychologie der Architektur*, if we were exclusively optical creatures, aesthetic judgment of the physical world would be precluded to us. Are the amazement and sense of elevation transmitted to us by the contemplation of a Gothic cathedral conceivable in purely algorithmic terms? Is it conceivable to divorce aesthetic experience from our daily muscular, tactile, and visceromotor experience of reality? Wölfflin (and together with him many others, among them Merleau-Ponty) maintained it was not, and we think that he was right.

We believe that our “natural” propensity to dualism is, on the one side, the product of our being asymmetrically positioned between mind and body, as [Helmuth Plessner](#) (2006) maintained. We are corporeal beings, but at the same time we

maintain that we *have* a body. On the other side, the account of the historical result of the progressive de-centralization of the anthropological dimension leads us to be dualist. We are no longer the living image of God, we are no longer at the centre of the universe, and perhaps in post-modern times we are no longer even subjects or selves. What are we left with but with the claim of the total otherness and discontinuity of our cognitive social life and its underlying processes? Their immaterial nature, or more exactly their total otherness in relation to a corporeity whose animal origin or essence is—evolutionarily speaking—pretty much clear, is perhaps the only way of reaffirming our uniqueness. The dualism between mind and body become, thus, a mechanism of defence. The so-called mental Rubicon that separates us from other non-human living beings is a very powerful anti-depressive argument for a disorientated humanity.

At this point, however, a clarification is required in order to avoid unpleasant misunderstandings. It is beyond doubt that the least intelligent among humans is incommensurably different and *other* in relation to the most intelligent among chimpanzees, despite their almost complete sharing of a genetic endowment. The point is that this quantum leap can be explained, perhaps, by remaining within an evolutionary framework that does not look for discontinuities founded on theories of “cognitive catastrophes,” genetic big bangs (as in the case of the so-called “grammar gene” invoked by Pinker), and so forth. The mysterious uniqueness—and loneliness—of humankind in the universe proves more comprehensible, or at least more easily approachable, if empirically investigated after having set aside the anti-continuist and self-consoling recipes of classic cognitive science. In our continuist approach, humankind is not special, because his evolution follows the same laws that regulate evolution of all other animals and is in continuity with evolutionary paths of other animals. However, our peculiar evolutionary path led us to acquire species-specific characteristics that only human beings share.

[Sigmund Freud](#) realized long before others how much the self is a bodily self (1923). Freud also helped us to understand how little we know

about who we are, particularly when aspiring to ground this knowledge solely on self-questioning rationality. What are the drives of which Freud spoke but a further manifestation of the double status of our flesh? We are *Körper* (objectual and represented body) and *Leib* (lived body), as Edmund Husserl maintained. Today cognitive neuroscience can shed new light on the *Leib* by investigating the *Körper*. The point is not to reduce the *Leib* to the *Körper*, but to understand that the empirical investigation of the *Körper* can tell us new things about the *Leib*.

9 Provisional conclusions

In this paper we addressed and discussed the notions of intersubjectivity and of the self as indissolubly intertwined outcomes of the bodily and symbolic dimensions. We proposed that embodied simulation seems to be able to naturalize the notion of paradigm, thus naturalizing one of the processes that makes language reflexivity possible, and thus contributing to “creating” the human being. Being a subject perhaps means being a body that learns to express itself and to express its world thanks to the paradigm—embodied simulation—that allows one to go beyond the body while remaining anchored to it. A new understanding of intersubjectivity can benefit from a bottom-up study and characterization of the non-declarative and non-metarepresentational aspects of social cognition (see Gallese 2003a, 2007).

One key issue of the new approach to intersubjectivity we proposed here is the investigation of the neural bases of our capacity to be attuned to the intentional relations of others. At a basic level, our interpersonal interactions do not make explicit use of propositional attitudes. This basic level consists of embodied simulation processes that enable the constitution of a shared meaningful interpersonal space. The shared intersubjective space in which we live from birth constitutes a substantial part of our semantic space. Self and other relate to each other because are opposite extensions of the same correlative and reversible we-centric space (Gallese 2003a). Observer and observed are part of a dynamic system governed by reversible

rules. By means of intentional attunement, “the other” is much more than a different representational system; it becomes a bodily self, like us.

This new epistemological approach to intersubjectivity has the merit of generating predictions about the intrinsic functional nature of our social cognitive operations, cutting across, and not being subordinated to a specific ontology of mind, like that purported by the classic cognitivist approach. Open questions that need to be further investigated in the future concern the biological mechanisms underlying our species-specific forms of self-knowledge and intersubjectivity. Language will have a special role in this investigation. To what extent and how are symbolic operations constrained by biological mechanisms? Is this connection between symbolic representations and bodily mechanisms that has been responsible for our specificity? These and other questions will be object of investigation in the next years

Acknowledgements

This work was supported by the EU Grant Towards an Embodied Science of InterSubjectivity (TESIS, FP7-PEOPLE-2010-ITN, 264828) and by the KOSMOS Fellowship from Humboldt University, Berlin to VG.

The authors wish to thank Anna Strasser, Michael Pauen and two anonymous reviewers for their most useful comments and criticisms on an earlier version of this article.

Although both authors discussed and designed the article together, sections 2, 3, 4, 5, and 8 were written by Vittorio Gallese, while sections 6, and 7 were written by Valentina Cuccio. Sections 1 and 9 were written jointly by both authors.

References

- Abe, K. & Watanabe, D. (2011). Songbirds possess the spontaneous ability to discriminate syntactic rules. *Nature Neuroscience*, 14 (8), 1067-1074. [10.1038/nn.2869](https://doi.org/10.1038/nn.2869)
- Adrian, E. D. & Matthews, R. (1927a). The action of light on the eye: Part I. The discharge of impulses in the optic nerve and its relation to the electric changes in the retina. *The Journal of Physiology*, 63 (4), 378-414.

- (1927b). The action of light on the eye: Part II. The processes involved in retinal excitation. *The Journal of Physiology*, 64 (3), 279-301.
- Agamben, G. (2008). *Signatura Rerum. Sul Metodo*. Torino, IT: Bollati-Boringhieri.
- Ammaniti, M. & Gallese, V. (2014). *The birth of inter-subjectivity: Psychodynamics, neurobiology, and the self*. New York, NY: Norton.
- Anderson, M. L. (2010). Neural reuse: A fundamental reorganizing principle of the brain. *Behavioral Brain Sciences*, 33 (4), 245-266. [10.1017/S0140525X10000853](https://doi.org/10.1017/S0140525X10000853)
- Aristotle, (2012). *The art of rhetoric*. London, UK: Harper Press.
- Arzy, S., Overney, L. S., Landis, T. & Blanke, O. (2006). Neural mechanisms of embodiment: Asomatognosia due to premotor cortex damage. *Archives of Neurology*, 63 (7), 1022-1025. [10.1001/archneur.63.7.1022](https://doi.org/10.1001/archneur.63.7.1022)
- Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology* (59), 617-645. [10.1146/annurev.psych.59.103006.093639](https://doi.org/10.1146/annurev.psych.59.103006.093639)
- Bermúdez, J. L. (2003). *Thinking without words*. New York, NY: Oxford University Press.
- Berti, A., Bottini, G., Gandola, M., Pia, L., Smania, N., Stracciari, A., Castiglioni, I., Vallar, G. & Paulesu, E. (2005). Shared cortical anatomy for motor awareness and motor control. *Science*, 309 (5733), 488-491. [10.1126/science.1110625](https://doi.org/10.1126/science.1110625)
- Bloomfield, T. C., Gentner, T. Q. & Margoliash, D. (2011). What birds have to say about language. *Nature Neuroscience*, 14 (8), 947-948. [10.1038/nn.2884](https://doi.org/10.1038/nn.2884)
- Boulenger, V., Shtyrov, Y. & Pulvermüller, F. (2012). When do you grasp the idea? MEG evidence for instantaneous idiom understanding. *NeuroImage*, 59 (4), 3502-3513. [10.1016/j.neuroimage](https://doi.org/10.1016/j.neuroimage)
- Butterfill, S. A. & Sinigaglia, C. (2014). Intention and motor representation in purposive action. *Philosophy and Phenomenological Research*, 88 (1), 119-145. [10.1111/j.1933-1592.2012.00604.x](https://doi.org/10.1111/j.1933-1592.2012.00604.x)
- Cermolacce, M., Naudin, J. & Parnas, J. (2007). The “minimal self” in psychopathology: Re-examining the self-disorders in the schizophrenia spectrum. *Consciousness and Cognition*, 16 (3), 703-714. [10.1016/j.concog.2007.05.013](https://doi.org/10.1016/j.concog.2007.05.013)
- Chittka, L. & Niven, J. (2009). Are bigger brains better? *Current Biology*, 19 (21), R995-R1008. [10.1016/j.cub.2009.08.023](https://doi.org/10.1016/j.cub.2009.08.023)
- Chomsky, N. (1966). *Cartesian linguistics: A chapter in the history of rationalist thought*. Lanham, MD: University Press of America.
- Christensen, K. R. (2009). Negative and affirmative sentences increase activation indifferent areas in the brain. *Journal of Neurolinguistics*, 22 (1), 1-17. [10.1016/j.jneuroling.2008.05.001](https://doi.org/10.1016/j.jneuroling.2008.05.001)
- Condillaco, E. B. (2001). *Essay on the origin of human knowledge*. Cambridge, UK: Cambridge University Press.
- Cosmides, L. & Tooby, J. (1997). The multimodular nature of human intelligence. In A. Schiebel & J. W. Schopf (Eds.) (p. Origin and Evolution of Intelligence). Toronto, Canada: Jones and Bartlett Publishers.
- Culioli, A. (1968). La formalisation en linguistique. *Cahiers Pour l'Analyse*, 9, 106-117.
- Dehaene, S. (2005). Evolution of human cortical circuits for reading and arithmetic: The neuronal recycling hypothesis. In S. Dehaene, J.-R. Duhamel, M. D. Hauser & G. Rizzolatti (Eds.) *From Monkey Brain to Human Brain. A Fyssen Foundation Symposium* (pp. 133-157). Cambridge, MA: MIT Press.
- Descartes, R. (1637). *Discours de la methode pour bien conduire sa raison, & chercher la verité dans les sciences: plus la dioptrique, les meteores, et la geometrie, qui sont des essais de cette methode*. Leiden: Jan Maire.
- (1642). *Meditationes de prima philosophia, in quibus Dei existentia & animae humanae a corpore distinctio demonstrantur: his adjunctae sunt variae objectiones doctorum virorum in istas de Deo & anima demonstrationes, cum responsionibus authoris. 2nd edition*. Amsterdam, NL: Elsevir.
- Ehrsson, H. H., Spence, C. & Passingham, R. E. (2004). That's my hand! Activity in premotor cortex reflects feeling of ownership of a limb. *Science*, 305 (5635), 875-877. [10.1126/science.1097011](https://doi.org/10.1126/science.1097011)
- Ferri, F., Frassinetti, F., Costantini, M. & Gallese, V. (2011). Motor simulation and the bodily self. *PLoS One*, 6 (3), e17927. [10.1371/journal.pone.0017927](https://doi.org/10.1371/journal.pone.0017927)
- Ferri, F., Frassinetti, F., Ardizzi, M., Costantini, M. & Gallese, V. (2012). A sensorimotor network for the bodily self. *Journal of Cognitive Neuroscience*, 24 (7), 1584-1595. [10.1162/jocn_a_00230](https://doi.org/10.1162/jocn_a_00230)
- Firpo, L. (1940). *Bibliografia degli scritti di Tommaso Campanella*. Turin, I: V. Bona.
- Freud, S. (1923). *The ego and the I*. London, UK: The Hogarth Press Ltd.
- Gallagher, S. (2005). *How the body shapes the mind*. Oxford, UK: Clarendon Press.
- Gallese, V. (2000). The inner sense of action: Agency and motor representations. *Journal of Consciousness Studies*, 7 (10), 23-40.

- (2003a). The manifold nature of interpersonal relations: The quest for a common mechanism. *Philosophical Transactions of the Royal Society of London B*, 358 (1431), 517-528. [10.1098/rstb.2002.1234](https://doi.org/10.1098/rstb.2002.1234)
- (2003b). A neuroscientific grasp of concepts: From control to representation. *Philosophical Transaction of the Royal Society of London B*, 358 (1435), 1231-1240. [10.1098/rstb.2003.1315](https://doi.org/10.1098/rstb.2003.1315)
- (2005). Embodied simulation: From neurons to phenomenal experience. *Phenomenology and the Cognitive Sciences*, 4 (1), 23-48. [10.1007/s11097-005-4737-z](https://doi.org/10.1007/s11097-005-4737-z)
- (2006). Intentional attunement: A neurophysiological perspective on social cognition and its disruption in autism. *Cognitive Brain Research*, 1079 (1), 15-24. [10.1016/j.brainres.2006.01.054](https://doi.org/10.1016/j.brainres.2006.01.054)
- (2007). Before and below theory of mind: Embodied simulation and the neural correlates of social cognition. *Philosophical Transactions of the Royal Society of London B*, 362 (1480), 659-669. [10.1098/rstb.2011.0029](https://doi.org/10.1098/rstb.2011.0029)
- (2008). Mirror neurons and the social nature of language: The neural exploitation hypothesis. *Social Neuroscience*, 3 (3-4), 317-333. [10.1080/17470910701563608](https://doi.org/10.1080/17470910701563608)
- (2009). Motor abstraction: A neuroscientific account of how action goals and intentions are mapped and understood. *Psychological Research*, 73 (4), 486-498. [10.1007/s00426-009-0232-4](https://doi.org/10.1007/s00426-009-0232-4)
- (2011). Neuroscience and phenomenology. *Phenomenology & Mind*, 1, 33-48.
- (2013). Corpo non mente. Le neuroscienze cognitive e la genesi di soggettività ed intersoggettività. *Educazione Sentimentale*, 20, 8-24. [10.3280/EDS2013-020002](https://doi.org/10.3280/EDS2013-020002)
- (2014). Bodily selves in relation: Embodied simulation as second-person perspective on intersubjectivity. *Philosophical Transaction of the Royal Society B*, 369 (1644), 20130177-20130177. [10.1098/rstb.2013.0177](https://doi.org/10.1098/rstb.2013.0177)
- Gallese, V., Fadiga, L., Fogassi, L. & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, 119 (2), 593-609. [10.1093/brain/119.2.593](https://doi.org/10.1093/brain/119.2.593)
- Gallese, V., Keysers, C. & Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends in Cognitive Sciences*, 8 (9), 396-403. [10.1016/j.tics.2004.07.002](https://doi.org/10.1016/j.tics.2004.07.002)
- Gallese, V. & Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in reason and language. *Cognitive Neuropsychology*, 22 (3), 455-479. [10.1080/02643290442000310](https://doi.org/10.1080/02643290442000310)
- Gallese, V. & Sinigaglia, C. (2010). The bodily self as power for action. *Neuropsychologia*, 48 (3), 746-755. [10.1016/j.neuropsychologia.2009.09.038](https://doi.org/10.1016/j.neuropsychologia.2009.09.038)
- (2011a). How the body in action shapes the self. *Journal of Consciousness Studies*, 18 (7-8), 117-143.
- (2011b). What is so special with embodied simulation. *Trends in Cognitive Sciences*, 15 (11), 512-519. [10.1016/j.tics.2011.09.003](https://doi.org/10.1016/j.tics.2011.09.003)
- Gentner, T. Q., Fenn, K. M., Margoliash, D. & Nusbaum, H. C. (2006). Recursive syntactic pattern learning by songbirds. *Nature*, 440, 1204-1207. [10.1038/nature04675](https://doi.org/10.1038/nature04675)
- Glenberg, A. M. & Gallese, V. (2012). Action-based language: A theory of language acquisition, comprehension, and production. *Cortex*, 48 (7), 905-922. [10.1016/j.cortex.2011.04.010](https://doi.org/10.1016/j.cortex.2011.04.010)
- Gould, S. J. & Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm. A critique of the adoptionist programme. *Proceedings of the Royal Society of London*, 205 (1161), 281-288. [10.1098/rspb.1979.0086](https://doi.org/10.1098/rspb.1979.0086)
- Gross, C. G. (2002). Genealogy of the 'Grandmother Cell'. *Neuroscientist*, 8 (5), 512-518. [10.1177/107385802237175](https://doi.org/10.1177/107385802237175)
- Gross, C. G., Rocha-Miranda, C. E. & Bender, D. B. (1972). Visual properties of neurons in inferotemporal cortex of the macaque. *Journal of Neurophysiology*, 35 (1), 96-111.
- Guan, C. Q., Meng, W., Yao, R. & Glenberg, A. M. (2013). The motor system contributes to comprehension of abstract language. *PLoS One*, 8 (9), e75183. [10.1371/journal.pone.0075183](https://doi.org/10.1371/journal.pone.0075183)
- Hauser, M. D., Chomsky, N. & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298 (5598), 1569-1579. [10.1126/science.298.5598.1569](https://doi.org/10.1126/science.298.5598.1569)
- Hinzen, W. & Sheehan, M. (2013). *The philosophy of universal grammar*. Oxford, UK: Oxford University Press.
- Husserl, E. (1973). Cartesianische Meditationen und Pariser Vorträge. In S. Strasser (Ed.) Den Haag, NL: Martinus Nijhoff.
- Kaup, B., Yaxley, R. H., Madden, C. J., Zwaan, R. A. & Lüdtke, J. (2006). Experiential simulations of negated text information. *Journal of Experimental Psychology*, 60 (7), 976-990. [10.1080/17470210600823512](https://doi.org/10.1080/17470210600823512)
- Kaup, B., Lüdtke, J. & Zwaan, R. A. (2007). Processing negated sentences with contradictory predicates: Is a door that is not open mentally closed? *Journal of Pragmatics*, 38, 1033-1050. [10.1016/j.pragma.2005.09.012](https://doi.org/10.1016/j.pragma.2005.09.012)

- Ku, S. P., Logothetis, N. K., Tolias, A. S. & Goense, J. (2011). fMRI of the face-processing network in the ventral temporal lobe of awake and anesthetized macaques. *Neuron*, 70 (2), 352-362. [10.1016/j.neuron.2011.02.048](https://doi.org/10.1016/j.neuron.2011.02.048).
- Kumar, U., Padakannaya, P., Mishra, R. K. & Khetrapal, C. L. (2013). Distinctive neural signatures for negative sentences in Hindi: An fMRI study. *Brain Imaging and Behaviour*, 7 (2), 91-101. [10.1007/s11682-012-9198-8](https://doi.org/10.1007/s11682-012-9198-8)
- Legrenzi, P. & Umiltà, C. A. (2011). *Neuromania: On the limits of brain science*. Oxford, UK: Oxford University Press.
- Leslie, A. M. (2005). Developmental parallels in understanding minds and bodies. *Trends in Cognitive Science*, 9 (10), 459-462. [10.1016/j.tics.2005.08.002](https://doi.org/10.1016/j.tics.2005.08.002)
- Liuzza, M. T., Candidi, M. & Aglioti, S. M. (2011). Do not resonate with actions: Sentence polarity modulates corticospinal excitability during action-related sentence reading. *PLoS ONE*, 6, e16855. [10.1371/journal.pone.0016855](https://doi.org/10.1371/journal.pone.0016855)
- Lo Piparo, F. (2003). *Aristotele e il linguaggio*. Roma, IT: Laterza.
- Margoliash, D. & Nusbaum, H. C. (2009). Language. The perspective from organismal biology. *Trends in Cognitive Sciences*, 13 (12), 505-510. [10.1016/j.tics.2009.10.003](https://doi.org/10.1016/j.tics.2009.10.003)
- Pauen, M. (2012). The second-person perspective. *Inquiry*, 55 (1), 33-49. [10.1080/0020174X.2012.643623](https://doi.org/10.1080/0020174X.2012.643623)
- Piazza, F. (2008). *La retorica di Aristotele. Introduzione alla lettura*. Roma, IT: Carocci.
- Pinker, S. (1994). *The language instinct*. New York, NY: Harper Collins.
- (1997). *How the mind works*. New York, NY: Norton.
- Plessner, H. (2006). *I gradi dell'organico e l'uomo*. Torino, IT: Bollati-Boringhieri.
- Pulvermüller, F. (2013). How neurons make meaning: Brain mechanisms for embodied and abstract-symbolic semantics. *Trends in Cognitive Science*, 458 (470), 1-17. [10.1016/j.tics.2013.06.004](https://doi.org/10.1016/j.tics.2013.06.004)
- Quiroga, R., Reddy, L., Kreiman, G., Koch, C. & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, 435, 1102-1107. [10.1038/nature03687](https://doi.org/10.1038/nature03687)
- Rizzolatti, G., Fadiga, L., Gallese, V. & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3 (2), 131-141.
- Rizzolatti, G. & Sinigaglia, C. (2010). The functional role of the parieto-frontal mirror circuit: Interpretations and misinterpretations. *Nature Reviews Neuroscience*, 11 (4), 264-274. [10.1038/nrn2805](https://doi.org/10.1038/nrn2805)
- Ross, W. D. (Ed.) (1978). *Aristotles prior and posterior analytics*. Oxford, UK: Clarendon Press.
- Saxe, R. (2006). Uniquely human social cognition. *Current Opinion in Neurobiology*, 16 (2), 235-249. [10.1016/j.conb.2006.03.001](https://doi.org/10.1016/j.conb.2006.03.001)
- Tettamanti, M., Buccino, G., Saccuman, M. C., Gallese, V., Danna, M., Scifo, P., Fazio, F., Rizzolatti, G., Cappa, S. F. & Perani, D. (2005). Listening to action-related sentences activates fronto-parietal motor circuits. *Journal of Cognitive Neuroscience*, 7 (2), 273-281. [10.1162/0898929053124965](https://doi.org/10.1162/0898929053124965)
- Tomasino, B., Weiss, P. H. & Fink, G. R. (2010). To move or not to move: Imperatives modulate action-related verb processing in the motor system. *Neuroscience*, 168, 246-258. [10.1016/j.neuroscience.2010.04.039](https://doi.org/10.1016/j.neuroscience.2010.04.039)
- Usener, H. (1887). *Epicurea*. Leipzig: Teubner. *Italian translation by Ilaria Ramelli, Epicurea: Testi di Epicuro e testimonianze epicuree nell'edizione di Hermann Usener*. Milan, IT: Bompiani, 2002.
- Varela, F. J. & Shear, J. (Eds.) (1999). *The view from within: First-person approaches to the study of consciousness*. Bowling Green, OH: Imprint Academic.
- Virno, P. (2003). *Quando il verbo si fa carne. Linguaggio e natura umana*. Torino, IT: Bollati Boringhieri.
- (2011). *E così via all'infinito. Logica e antropologia*. Torino, IT: Bollati Boringhieri.
- Vogeley, K., May, M., Ritzl, A., Falkai, P., Zilles, K. & Fink, G. R. (2004). Neural correlates of first-person perspective as one constituent of human self-consciousness. *Journal of Cognitive Neuroscience*, 16 (5), 817-827. [10.1162/089892904970799](https://doi.org/10.1162/089892904970799)
- Vogeley, K. & Fink, G. R. (2003). Neural correlates of the first-person-perspective. *Trends in Cognitive Sciences*, 7 (1), 38-42.
- Wojciehowski, H. C. & Gallese, V. (2011). How stories make us feel. Toward an embodied narratology. *California Italian Studies*, 2 (1).
- Wölfflin, H. (1886). *Prolegomena zu einer Psychologie der Architektur*. München, GER: University of Munich.
- Zeki, S. (1993). *A vision of the brain*. London, UK: Wiley-Blackwell.

Multisensory Spatial Mechanisms of the Bodily Self and Social Cognition

A Commentary on Vittorio Gallese & Valentina Cuccio

Christian Pfeiffer

This commentary aims to find the right description of the pre-reflective brain mechanisms underlying our phenomenal experience of being a subject bound to a physical body (bodily self) and basic cognitive, perceptual, and subjective aspects related to interaction with other individuals (social cognition). I will focus on the proposal by Gallese and Cuccio that embodied simulation, in terms of motor resonance, is the primary brain mechanism underlying the pre-reflective aspects of social cognition and the bodily self. I will argue that this proposal is too narrow to serve a unified theory of the neurobiological mechanisms of both target phenomena. I support this criticism with theoretical considerations and empirical evidence suggesting that multisensory spatial processing, which is distinct from but a pre-requisite of motor resonance, substantially contributes to the bodily self and social cognition.

My commentary is structured in three sections. The first section addresses social cognition and compares embodied simulation to an alternative account, namely the attention schema theory. According to this theory we pre-reflectively empathize with others by predicting their current state of attention which involves predicting the spatial focus of attention. Thereby we derive a representational model of their state of mind. On this account, spatial coding of attention, rather than motor resonance, is the primary mechanism underlying social cognition. I take this as a theoretical alternative complementing motor resonance mechanisms.

The second section focuses on the bodily self. Comparison of the brain networks of the bodily self and social cognition reveals strong overlap, suggesting that both phenomena depend on shared multisensory and sensorimotor mechanisms. I will review recent empirical data about altered states of the bodily self in terms of self-location and the first-person perspective. These spatial aspects of the bodily self are encoded in brain regions distinct from the brain network of embodied simulation. I argue that while motor resonance might contribute to body ownership and agency, it does not account for spatial aspects of the bodily self. Thus, embodied simulation appears to be a necessary but insufficiently “primary” brain mechanism of the bodily self and social cognition.

The third section discusses the contributions of the vestibular system, i.e., the sensory system encoding head motion and gravity, to the bodily self and social cognition. Vestibular cortical processing seems relevant to both processes, because it directly encodes the world-centered direction of gravity and allows us to distinguish between motions of the own body and motions of other individuals and the external world. Furthermore, the vestibular cortical network largely overlaps with those neural networks relevant to the bodily self and social cognition. Thus, the vestibular system may play a crucial role in multisensory spatial coding relating the bodily self to other individuals in the external world.

Keywords

Attention schema | Bodily self | Embodied simulation | First-person perspective | Mirror neurons | Self-location | Social cognition | Vestibular system

Commentator

[Christian Pfeiffer](#)

christian.pfeiffer@epfl.ch

Ecole Polytechnique Fédérale
Lausanne, Switzerland

Target Authors

[Vittorio Gallese](#)

vittorio.gallese@unipr.it

Università degli Studi di Parma
Parma, Italy

[Valentina Cuccio](#)

valentina.cuccio@unipa.it

Università degli Studi di Palermo
Palermo, Italy

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

The paper by Gallese and Cuccio provides an integrated theoretical framework explaining how the brain and body relate to social cognition, the human self, and language. The authors review empirical evidence from electrophysiological and neuroimaging studies supporting embodied simulation (ES) theory (Gallese & Cuccio [this collection](#), p. 8). According to ES, the brain covertly simulates the bodily actions, perceptions, and emotions observed in other individuals by using parts of our neural architecture involved in acting, sensing, and feeling emotions. Thereby, we infer the goals, intentions, and states of mind of others in a pre-reflective and non-conceptual fashion. But the authors take this a step further and propose that ES is the key mechanism underlying, and hence unifying, both social cognition, the human self, and language. Throughout the paper, the authors emphasize the tight functional coupling between the body and the brain, which when taken into account bears the potential to significantly advance the scientific study of the hard problem of consciousness (Chalmers 1996).

This commentary on Gallese and Cuccio aims to find the right description of the brain mechanisms underlying pre-reflective aspects of both the bodily self and social cognition. Specifically, I will focus on Gallese and Cuccio's central claim that ES, based on motor resonance and neural processing in the motor system, is the primary brain mechanism underlying pre-reflective representations of the bodily self and social cognition (Gallese & Cuccio [this collection](#), pp. 8–14). I ask the following questions: Could there be an alternative theory or empirical evidence countering the claim of a primacy of motor resonance underlying social cognition and the bodily self? Which brain mechanisms in addition to motor resonance might contribute to pre-reflective aspects of social cognition and the bodily self? I will defend the following three theses:

(1) Social cognition and the bodily self depend on multisensory spatial coding, which is distinct from motor resonance.

Thus, motor resonance may be a necessary but insufficiently “primary” brain mechanism of social cognition and the bodily self (cf. section 1, 2).

(2) The brain networks underlying social cognition and the bodily self largely overlap. Specific functional associations exist (a) between motor resonance and body ownership/agency and (b) between multisensory spatial coding and self-location/the first-person perspective (cf. section 2).

(3) The vestibular system, i.e., the sensory system encoding head motion and gravity, might provide unique information used for multisensory spatial coding that relates the bodily self to other individuals and the external world. This is further suggested by the large overlap existing between the human vestibular cortex and the brain networks underlying the bodily self and social cognition (cf. section 3).

My commentary is structured in three sections. In the first section I shall compare ES to an alternative theory of social cognition that assigns priority to spatial coding of attention, rather than to motor resonance. I shall show that both theories bear the potential that their proposed brain mechanisms cooperatively work together in order to support social cognition. The second section addresses the bodily self. I shall review data from neurological patients and full-body illusion experiments, which highlight the importance of two spatial aspects of the bodily self not mentioned by Gallese and Cuccio, i.e., self-location and the first-person perspective. These spatial aspects of the bodily self depend primarily on multisensory integration and on cortical processing outside regions involved in ES. Furthermore, comparisons between the brain networks encoding the bodily self and social cognition show large overlaps, suggesting shared functional mechanisms. In the third section I propose that because multisensory spatial processing appears to be critical for

the bodily self and social cognition, important contributions may come from the vestibular system (Lenggenhager & Lopez [this collection](#)). I shall show that the vestibular cortical network largely overlaps with the brain networks underlying the bodily self and social cognition. I shall discuss potential contributions of vestibular cortical processing to these target phenomena and suggest directions for future research.

2 Is social cognition based on motor resonance or attention tracking?

Social cognition refers to cognitive processes, perceptions, and subjective experiences related to interaction with conspecifics. This section asks: Which are the brain mechanisms underlying pre-reflective aspects of social cognition? Could there be alternative theories and empirical evidence countering the primary role of motor resonance?

Gallese and Cuccio propose that social cognition mainly depends on ES based on motor resonance and processing of mirror neurons (see citations in [Gallese & Cuccio this collection](#)). Mirror neurons were initially discovered in fronto-parietal networks of the macaque monkey brain. They are a specific type of canonical neuron involved in planning and executing hand actions and were found to be activated both when the monkey executed a specific grasping or reaching action and when the monkey passively observed somebody performing similar actions ([Gallese et al. 1996](#); [Rizzolatti et al. 1996](#)). Neuroimaging studies in humans also showed mirror neuron-like activation patterns at the level of populations of neurons in distinct brain regions—mainly the ventral premotor cortex (vPM), the intraparietal sulcus (IPS), but also the insula cortex and the secondary somatosensory cortex ([Rizzolatti & Sinigaglia 2010](#); see also figure 1a gray dots). ES proposes that based on mirror neurons the brain maps observed actions into an action space, into motor potentialities, within our hierarchically-organized motor system, and thereby infers and predicts the action goals of the individual. In this way it penetrates the state of mind of the other, and thus links self and other in a pre-reflective

empathical fashion ([Gallese & Cuccio this collection](#), p. 7).

I would like to point out that motor resonance, i.e., the mapping of observed actions into motor potentialities, necessarily depends on multisensory spatial coding. I argue that this is the case because of five points: First, the brain has access to the physical world only through the different sensory receptors of the body that bombard it with exteroceptive (e.g., vision, audition), proprioceptive (somatosensory, vestibular), and interoceptive (somatosensory, visceral) signals. Second, these multisensory signals must be integrated according to their spatial and temporal parameters ([Stein & Stanford 2008](#)) to inform neural representations of the states of the body and of the world around us—including the agents whose actions are subject to motor resonance. Third, the observed movements of these agents are coded in coordinates distinct from the egocentric spatial frame of reference upon which our motor system operates. Fourth, the brain must necessarily perform spatial transformations of the observed movements by the other agent into the egocentric frame of reference, upon which motor resonance can operate. In sum, multisensory spatial coding is a pre-requisite of motor resonance.

According to Gallese and Cuccio, the outcomes of such multisensory spatial coding are readily available to the brain network of ES through anatomical connections to the vPM that are “anatomically connected to visual and somatosensory areas in the posterior parietal cortex and to frontal motor areas” ([Gallese & Cuccio this collection](#), p. 10). However, it seems that the multisensory spatial coding required for a precise description of complex motor acts might be computationally costly. Might there be a computationally more effective alternative by which multisensory spatial coding is used to decode the intentions of observed agents?

The attention schema (AS) theory of awareness ([Graziano 2013](#); [Graziano & Kastner 2011](#)) proposes that brain mechanisms related to attention and spatial coding, which are distinct from neural processing relevant to ES, primarily underlie pre-reflective aspects of social cognition. Graziano and Kastner define *atten-*

tion as an information-handling mechanism of the brain that serves to give priority to some information (e.g., representational features) out of several equally probable alternatives that are in constant competition for awareness. Furthermore, *awareness* is defined as the process of consciously experiencing something, it is the process of relating the subject (i.e., a phenomenal self, see also Metzinger 2003) to the object/content of experience. Graziano and Kastner summarize AS as follows:

[Awareness is information and] depends on some system in the brain that must have computed [it] [...]; otherwise, the information would be unavailable for report. [...] People routinely compute the state of awareness of other people [and] the awareness we attribute to another person is our reconstruction of that person's attention. [...] The same machinery that computes socially relevant information [...] also computes [...] information about our own awareness. [...] Awareness is [...] a perceptual model [...] a rich informational model that includes, among other computed properties, a spatial structure. [...] Through the use of the social perceptual machinery, we assign the property of awareness to a location within ourselves. (Graziano & Kastner 2011, pp. 98–99)

Related to social cognition, AS proposes that by using a schematic representation of the state of attention of other individuals—including a prediction of the spatial location of their focus of attention—we predict the current state of awareness of the individual, which is informative about their intentions and potential future actions. In short: Awareness of others is an attention schema. As compared to ES, AS is a relatively recent theory that requires extensive empirical studies. Yet the evidence so far shows that indeed the brain has a neural circuitry for monitoring the spatial configuration of one's own attention independent of the sensory modality (Downar et al. 2000), including the direction of gaze (Beck & Kastner 2009; Desimone &

Duncan 1995). These structures are the proposed neural expert system upon which AS is based and consist of the right-hemispheric temporo-parietal junction (TPJ) and superior temporal sulcus (STS) (see figure 1a in black). Notably, this expert system relevant to AS shows little anatomical overlap with the neural structures relevant to ES (figure 1a compare black with gray).

Because the AS relies on coding of the spatial relationship between the location of the observed individual and the likely spatial location of this individual's attention (i.e., independent of a particular sensory modality), the required spatial computations seem simple and straightforward. They require two points, i.e., the individual as a reference point and the potential spatial location of the attention of that individual. According to AS, using such spatial labeling the brain is able to simultaneously track the aware and attending minds of several individuals simultaneously. Thus, spatial coding in the context of AS appears to be less complex and less computationally demanding than spatial transformations underlying ES (see above).

Which of these seemingly distinct brain mechanisms proposed by AS and ES more plausibly underlies social cognition: the neural expert system decoding the state of attention according to AS or the mirror mechanism system decoding observed motor plans according to ES? It has been proposed that AS and ES may in principle work together. Graziano and Kastner propose that the expert system of AS may take a leading role by formulating a hypothesis about the state of awareness of an individual that is likely to drive further behavior and therefore provide a set of predictions based upon which motor resonance could more efficiently perform simulations (Graziano & Kastner 2011). Motor resonance would thus add richer detail to the state-of-attention hypothesis made by the expert system.

This combined mechanism is compatible with the predictive processing principle (Clark this collection; Hohwy 2013, this collection), which has been proposed relevant to the bodily self (Apps & Tsakiris 2013; Limanowski &

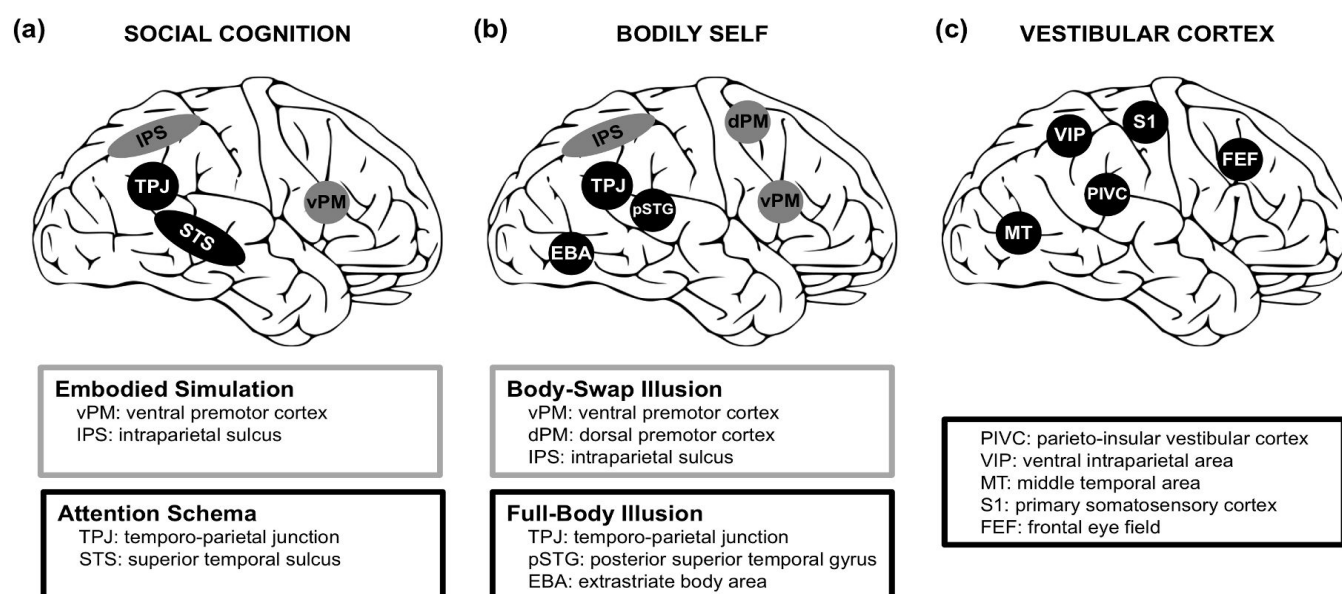


Figure 1: Summary of cortical brain regions involved in social cognition, the bodily self, and vestibular processing. (a) Whereas for social cognition there is little overlap between the brain regions proposed relevant for the attention schema (*in black*) and embodied simulation (*in gray*), both sets of brain regions overlap with (b) the brain network of the bodily self as identified by full-body illusion experiments manipulating self-location and first-person perspective (*in black*) and the body-swap illusion manipulating mainly body ownership (*in gray*). (c) The human vestibular cortical regions (*in black*) are widely distributed and overlap with several regions relevant to both the bodily self and social cognition. (The images are derived from images by NASA, licensed under creative commons.)

Blankenburg 2013; Seth this collection). According to *predictive processing* the brain constantly predicts the potential causes of sensory input by minimizing prediction errors via update of the predicted causes or by action that changes sensory input (Friston 2005). Applying the predictive processing principle to Graziano and Kastner's proposal that AS is a hypothesis-generating tool to which ES adds further detail, one could conceive of both mechanisms as different predictive processing modules aimed at anticipating the state of awareness and of intentional actions observed in others. Although no empirical study so far has addressed this specific hypothesis, a recent functional magnetic resonance imaging study found that predictive processing principles accounted for the blood oxygen-level dependent activity related to the perception of faces, which is an important perceptual function for social cognition in the human species (Apps & Tsakiris 2013).

These common and distinct predictions based on ES, AS, and predictive processing

call for empirical research aimed at providing evidence to further refine, integrate, or reject them.

3 Multisensory and motor mechanisms of the multifaceted bodily self

The *bodily self* refers to the phenomenal experience of being an experiencing subject (i.e., a phenomenal self) bound to a physical body, which gives rise to the dual nature of the body (Husserl 1950; Gallese & Cuccio this collection, p. 2). The unified experience of being a bodily self can be decomposed into different aspects, including the experience that we identify with a particular body (self-identification or *body ownership*), the experience that the self is situated in a specific spatial location (*self-location*), that we take a specific experiential perspective at the world (*first-person perspective*), and that we are the authors of our actions, including having control of attentional focus (*agency*; (Blanke 2012; Ehrsson 2012; Jeannerod 2003; Metzinger 2003).

In their paper, [Gallese & Cuccio](#) highlight the relevance of mirror mechanisms, in particular related to processing in the cortical motor system, to the sense of body ownership and the sense of agency, in particular in the context of action and action observation:

This minimal notion of the self, namely the bodily self as power-for-action [...], tacitly presupposes ownership of an action-capable agentive entity; hence, it primarily rests upon the functionality of the motor system. ([this collection](#), p. 10)

However, recent philosophy of mind and cognitive neuroscience research reveals the crucial role of *spatial aspects of the bodily self*, consisting of a first-person perspective and self-location. In this section I shall compare the brain network contributing to spatial aspects of the bodily self with the brain network underlying body ownership and ask: Do these neuroimaging results support the proposal that motor resonance is a primary mechanism underlying all aspects of the bodily self? What is the relationship between the neural networks of the bodily self and social cognition? Which functional associations can be derived from this?

3.1 Brain mechanisms of spatial aspects of the bodily self

The phenomenal experience of being a subject is associated with a spatial location, which typically is the space of the physical body (see also [Alsmith & Longo 2014](#); [Limanowski & Hecht 2011](#)). However, there are exceptions to these prototypical states of the bodily self in neurological disorders and experimental illusions pointing to a specific set of brain regions involved in spatially linking the phenomenal self to the physical body.

Which brain mechanisms link the phenomenal self to the physical body to give rise to the dual nature of the body as lived body and as physical object? Research in neurological patients who have had out-of-body experiences (OBE) shows that damage or interference with the right TPJ can lead to dissociations between

the bodily self and physical body ([Blanke et al. 2004](#); [Blanke et al. 2002](#); [De Ridder et al. 2007](#); [Ionta et al. 2011](#)). During an OBE, patients typically experience a disembodied self-location in elevation above their physical body, and an altered first-person perspective that originates from an elevated location in the room and is directed downwards at the physical body ([Blanke et al. 2004](#); [Metzinger 2009](#)). These patients do not identify with their physical body but with an illusory double outside of the borders of the physical body. At the phenomenological level, self-location and the first-person perspective are often experienced as having their spatial origin in the same position. However, during OBE there are instances where self-location can be dissociated from the first-person perspective in different sensory modalities ([De Ridder et al. 2007](#)). Further evidence from asomatic OBEs and bodiless dreams suggests that a phenomenal first-person perspective may be reducible to a single point in space ([Windt 2010](#)). In fact, vestibular hallucinations systematically preceded OBEs in patients with sleep paralysis, i.e., a motor paralysis characterised by the transient inability to execute bodily actions when waking up from sleep ([Cheyne & Girard 2009](#)), showing further dissociations of the spatial location of the bodily self and the physical body and links to sensory processing. These studies seem to suggest that the first-person perspective and self-location may depend on different neural mechanisms ([Blanke 2012](#)).

OBE in epileptic patients can be induced by subcortical electrical stimulation of a specific intensity at the TPJ. However, stimulating the same brain region with either lower or higher stimulation intensity induces bodily sensations (including vestibular, visual, somatosensory, kinesthetic sensations) without inducing an OBE ([Blanke et al. 2002](#)). These observations gave rise to the idea that the spatial aspects of the bodily self are based on the accurate integration of multisensory signals (i.e., which was perturbed by electrical stimulation in the patient in [Blanke et al. 2002](#), which are sensory signals from personal space to sensory signals from the external environment [Blanke et al. 2004](#)).

These clinical observations in patients were corroborated by different full-body illusion experiments in healthy subjects, such as the so-called “body-swap illusion” (Petkova & Ehrsson 2008; Petkova et al. 2011; van der Hoort et al. 2011), the “full-body illusion” (Ionta et al. 2011; Lenggenhager et al. 2009; Lenggenhager 2007; Pfeiffer et al. 2013; Pfeiffer, Schmutz & Blanke 2014), and the “out-of-body illusion” (Ehrsson 2007; Guterstam & Ehrsson 2012). In these experiments, healthy subjects receive conflicting signals about the spatial location of their body and of the temporal synchrony of exteroceptive and interoceptive signals, including somatosensory, cardiac, and vestibular signals that at the same time are applied to a virtual or fake body seen by the subject (Aspell et al. 2013; Ionta et al. 2011; Pfeiffer et al. 2013; Pfeiffer et al. 2014). For example, in the *full-body illusion*, synchronous stroking of a virtual or fake body seen from a distance can induce the feeling in participants that they are more closely located to the position of the virtual or fake body, and that they experience and increase of ownership for the seen body. The brain regions involved in these spatial experimental manipulations of the experienced bodily self most consistently involve the right TPJ region, but also draw on somatosensory and visual regions that process the sensory inputs (Blanke 2012; Ionta et al. 2011; figure 1b in black). Recently, several studies have manipulated visual and vestibular signals about the direction of gravity, affecting self-location and perspective and thus showing that those visual spatial cues affect our subjective experience of the first-person perspective (Ionta et al. 2011; Pfeiffer et al. 2013). These authors presented images on virtual-reality goggles showing visual gravitational cues, similar to the visual perspective during an OBE showing a scene from an elevated spatial location and a visual viewpoint directed downwards into the room. At the same time the somatosensory and the vestibular signals received by the participant, who was lying on the back, suggested that the physical body was oriented upwards with respect to veridical gravity. Thus the visual gravity cues (i.e., downwards) and the vestibular gravity cues (i.e., upwards) were in directional

conflict. When the full-body illusion was induced under these conflicting conditions, participants reported subjective changes in their experienced direction of the first-person perspective (upward or downward) in line with experimentally-induced multisensory conflict (Ionta et al. 2011; Pfeiffer et al. 2013).

3.2 Brain mechanisms of body ownership

A different brain network encodes experimental manipulations of another aspect of the bodily self: body ownership. This was shown by the *body-swap illusion* (Petkova & Ehrsson 2008; Petkova et al. 2011), during which the participant views from a first-person visual viewpoint the body of a mannequin or another person. Thus no conflict between the visual spatial coordinates of the participant’s physical body and the visually-perceived location of the mannequin is presented. However, conflicting sensory information about the shape, gender, size, or overall spatial context surrounding the virtual body were presented that typically prevented feeling ownership of the virtual body. If under these conditions visuo-tactile stroking on the abdomen of the participant and the virtual body was synchronously administered, an illusion of ownership for the body emerged, reflected in increased responses to threatening the mannequin. In different variants of the body-swap illusion subjects reported experiencing and adopting different sizes of both the virtual body and the contextual environment (Petkova & Ehrsson 2008; Petkova et al. 2011; van der Hoort et al. 2011). Neuroimaging experiments of the body-swap illusion show activation of the vPM and IPS regions, notably without involving actions made by subjects or performed by the virtual body (Petkova et al. 2011). These brain regions are key nodes of the mirror mechanism network of ES (see Serino et al. 2013). For a recent review see figure 1b.

3.3 A shared brain network of bodily self and social cognition

Although the neuroimaging evidence so far suggests that distinct brain regions encode the spa-

tial aspects of the bodily self and body ownership (Blanke 2012; Serino et al. 2013), the ensemble of those bodily self-encoding regions closely matches the brain regions relevant for social cognition (compare in figure 1a with figure 1b). These empirical data indeed suggest that the bodily self and social cognition are encoded by at least overlapping neural circuits supporting the proposal of ES that neural capacities to control and monitor the own body are used in understanding others.

These neuroimaging data suggest particular functional associations between different aspects of social cognition and the bodily self. In particular, the brain network of ES anatomically overlaps with regions encoding experimentally-induced changes in body ownership during the body-swap illusion (figure 1a–b in gray), which involves spatial congruence of the observational viewpoint and position of the fake body and the participant’s body. A second association can be observed between the brain network of AS and the brain regions encoding spatial aspects of the bodily self, as manipulated during the full-body illusion (figure 1a–b in black). During the latter, the position and observational viewpoints of the virtual body and the participant’s body are in spatial conflict, and thus closely resemble social interaction settings.

Based on these functional and neuroanatomical observations, I propose that ES seems to contribute to the bodily self and social cognition in a way primarily related to the sense of body ownership and agency. However, ES does not account for multisensory spatial representations that relate the physical body to the bodily self in space. These spatial aspects of the bodily self are encoded by brain regions outside of the brain network of ES, and rather resemble those brain regions relevant for coding the spatial configuration of attention (or awareness, according to AS).

Because two crucial aspects of the bodily self, i.e., self-location and the first-person perspective, are encoded in the TPJ region, and full-body illusions show that they can be manipulated without action or motor manipulations, it seems implausible that ES as based on

motor resonance is the primary brain mechanism underlying the bodily self. Instead, the brain networks coding self-location and the first-person perspective, which overlap with brain regions proposed to encode spatial aspects of an attention schema (see figure 1), seem to contribute to at least an equal degree to both the bodily self and social cognition. Thus, ES seems to be a necessary but insufficiently “primary” brain mechanism underlying the bodily self and social cognition.

I do not mean to imply that these are independent processes, because it is possible that they cooperatively work together (Graziano & Kastner 2011). However, I think that Gallese and Cuccio’s claim of a primacy of motor resonance underlying the multifaceted aspects of the bodily self and social cognition is questionable on empirical and theoretical grounds.

4 Vestibular contributions to the bodily self and social cognition

In the previous sections I have provided theoretical considerations and empirical evidence assigning a critical role to multisensory spatial processing in the neural computations underlying representations of the bodily self and social cognition. This section will further examine the multisensory mechanisms relating the space of the bodily self to other individuals and the external world. I propose that important contributions to the brain’s multisensory spatial coding might come from a particular sensory system, i.e., the vestibular system, which has often been neglected in studies of higher brain functions related to subjectivity and intersubjectivity. I will ask: What might be the functional contribution of the vestibular system to pre-reflective representations of the bodily self and social cognition? How does the human vestibular cortex relate to the neural networks of the bodily self and social cognition?

The *vestibular system* consists of sensory organs in the inner ear that sense accelerations of the head in space, including rotational and linear movement of the head and whole body and the constant acceleration of gravity on earth (Day & Fitzpatrick 2005). Vestibular sig-

nals are processed by subcortical and cortical structures (Angelaki & Cullen 2008; Cullen 2012; Lopez & Blanke 2011). Research initially focused on subcortical processing as related to gaze control, postural stabilization, and neural computations of head motion directions (Fernandez & Goldberg 1971; Goldberg & Fernandez 1971). More recently, studies have revealed the contribution of vestibular cortical processing to spatial cognition, body perception, and the bodily self (see Lenggenhager & Lopez this collection; Lopez & Blanke 2011; Pfeiffer et al. 2014 for reviews). These studies show that vestibular cortical processing is based on a neural network of distinct, distributed, and multisensory cortical regions. In distinction from any other sensory modality, there is no primary vestibular cortex that processes purely vestibular signals. Instead, a core vestibular cortical input region, the human parieto-insular vestibular cortex (PIVC; Lopez et al. 2012; zu Eulenburg et al. 2012), processes vestibular, somatosensory, and visual signals and is connected to a number of multisensory brain regions in the parietal, temporal, cingulate, and frontal regions (figure 1c).

The vestibular system contributes to spatial aspects of the bodily self. For instance, OBEs were associated with vestibular sensation, such as floating in elevation (Blanke et al. 2004; Blanke & Mohr 2005; Blanke et al. 2002), and vestibular sensations preceded OBEs in persons with sleep paralysis (Cheyne & Girard 2009). Other studies presented conflicting visual and vestibular signals about earth gravity during the full-body illusion and induced changes in the subjectively-experienced spatial direction of the first-person perspective and self-location (Ionta et al. 2011; Pfeiffer et al. 2013). Thus, it has been argued that vestibular cortical processing does not merely signal the motions of the own body and the external world, but is also constitutive of spatial aspects of the bodily self (Lopez et al. 2008; Pfeiffer et al. 2014).

Previously, Lopez et al. (2013), Deroualle & Lopez (2014), and Lenggenhager & Lopez (this collection) have argued that the vestibular system probably contributes to social cognition. I will briefly summarize their main ar-

guments and complement them with own points:

First, because the human species evolved under the steady influence of the earth's gravitational field, adaptation to gravity also framed and affected action, perception, and social interaction. More recently, research has shown that the brain hosts internal models of gravity, representing the effects of gravity on the motion of objects under the influence of gravity, of self-motion, of bodily actions, and of the direction of the gravitational acceleration. Those internal models of gravity strongly overlap with the vestibular cortex (Indovina et al. 2005; Indovina et al. 2013; McIntyre et al. 2001; Sciutti et al. 2012). More evidence for a vestibular contribution to social perception comes from studies showing the effects of gravitational signals on the perception of emotional faces (Thompson 1980) and the perception of the spatial orientation of bodies (Lopez et al. 2009).

Second, the vestibular system might contribute to social cognition because it detects head motions in space and hence directly enables us, when compared to other sensory signals, to discern movements made by our own body from motions of other individuals and motions of the external environment (Deroualle & Lopez 2014).

Third, mental spatial transformation of the own visual viewpoint to that of another person presents an important underlying cognitive aspect of social cognition (Furlanetto 2013; Hamilton 2009; Newen & Vogeley 2003; also cited by Gallese & Cuccio this collection, pp. 9–11). More direct evidence supporting this hypothesis comes from a recent study that showed that physical whole-body rotations, which stimulate the vestibular sensory organs, affected the ability of participants to perform mental spatial transformations (van Elk & Blanke 2013).

Fourth, I have argued in previous sections of this commentary that multisensory spatial coding is a critical prerequisite that underlies pre-reflective brain mechanisms of the bodily self and social cognition. Because the vestibular cortical processing has been strongly associated with multisensory integration (for review see Lopez & Blanke 2011), it is likely that vestibular

lar signals shape multisensory spatial coding relevant to the bodily self and social cognition (Deroualle & Lopez 2014; Pfeiffer et al. 2014).

Fifth, the distributed multisensory vestibular cortical network clearly overlaps with the neural structures involved in social cognition and the bodily self, which suggests that there is a functional contribution on the part of vestibular processing to these phenomena (compare figure 1c to 1a and 1b; compare also to Deroualle & Lopez 2014).

Together, these five points suggest that the vestibular system may be a promising candidate for future studies of the sensorimotor mechanisms of social cognition, which should motivate research on the intersection of vestibular cortical processing, mirror mechanisms, and intersubjectivity. These studies may, for instance, question how vestibular stimulation affects our ability to reconstruct the process of attention of another person, a function critical in the AS framework. Although the vestibular system is related to reflexive motor control, it is not clear whether it also affects motor resonance (see Deroualle & Lopez 2014 for a related proposal). One might ask whether vestibular processing facilitates or inhibits motor resonance and our understanding of intentional action observed in others. How about vestibular contributions to theory of mind and reasoning? On the other hand, does social interaction modulate vestibular functions, such as self-motion perception, postural stabilization, and gaze control? These questions address the role of vestibular processing in functional mechanisms relevant to the AS and ES frameworks. Furthermore, empirical research addressing the causal relationship between the AS and ES brain mechanisms and the bodily self and social cognition are needed, for instance by brain lesion analysis or direct brain stimulation.

5 Conclusion

At the beginning of this paper I asked which brain mechanisms underlie pre-reflective representations of the bodily self and social cognition. ES, based on motor resonance, substantially contributes to the representation of the bodily

self and social cognition. However, a unified theory of the neural basis of these target phenomena cannot assign a primary role to motor resonance. I have argued that multisensory spatial coding is at least of equal importance and probably more basic than ES in contributing to several key aspects of the bodily self and social cognition.

Specifically, I have argued that:

- (1) Social cognition and the bodily self depend on multisensory spatial coding, which is distinct from motor resonance. Thus, motor resonance may be a necessary but insufficiently “primary” brain mechanism of social cognition and the bodily self (cf. section 1, 2).
- (2) The brain networks underlying social cognition and the bodily self largely overlap. Specific functional associations exist (a) between motor resonance and body ownership/agency and (b) between multisensory spatial coding and self-location/the first-person perspective (cf. section 2).
- (3) The vestibular system, i.e., the sensory system encoding head motion and gravity, might provide unique information used for multisensory spatial coding that relates the bodily self to other individuals and the external world. This is further suggested by the large overlap existing between the human vestibular cortex and the brain networks underlying the bodily self and social cognition (cf. section 3).

A unifying theory of pre-reflective brain mechanisms of the bodily self and social cognition must be able to account for the empirical evidence reviewed here; and it seems that such a theory cannot exclusively depend on motor resonance. Multisensory spatial coding, motor mechanisms, but also representations of the process of attention appear highly relevant to bodily self and social cognition.

I agree with Gallese & Cuccio (this collection, pp. 3–7) that cognitive neuroscience cannot fully explore these exciting topics by limit-

ing itself to a specific neuroimaging method, such as functional magnetic resonance imaging. Instead, we should exploit multi-method approaches in search for correlative and causal evidence relating brain function and anatomy to the phenomenology of the bodily self and social cognition. The body, but also the spatial representation of the world around us, are relevant to understanding brain function, and when taken into account can lead to novel approaches to phenomenal analysis of subjective experience. But we should be careful in assigning priority to a single brain mechanism when aiming to explain the human self and intersubjectivity. Scrutiny and dialogue at the intersection of philosophy of mind and cognitive neuroscience are necessary in order to advance our understanding of the nature of the human mind.

Acknowledgments

I thank Thomas Metzinger, Jennifer Windt, and two anonymous reviewers for constructive comments on an earlier version of this commentary.

References

- Alsmith, A. J. & Longo, M. R. (2014). Where exactly am I? Self-location judgements distribute between head and torso. *Consciousness and Cognition*, 24, 70-74. [10.1016/j.concog.2013.12.005](https://doi.org/10.1016/j.concog.2013.12.005)
- Angelaki, D. E. & Cullen, K. E. (2008). Vestibular system: the many facets of a multimodal sense. *Annual Reviews in Neuroscience*, 31, 125-150. [10.1146/annurev.neuro.31.060407.125555](https://doi.org/10.1146/annurev.neuro.31.060407.125555)
- Apps, M. A. & Tsakiris, M. (2013). Predictive codes of familiarity and context during the perceptual learning of facial identities. *Nature Communications*, 4 (2698), 2698-2698. [10.1038/ncomms3698](https://doi.org/10.1038/ncomms3698)
- Aspell, J. E., Heydrich, L., Marillier, G., Lavanchy, T., Herbelin, B. & Blanke, O. (2013). Turning body and self inside out: Visualized heartbeats alter bodily self-consciousness and tactile perception. *Psychological Science*, 24 (12). [10.1177/0956797613498395](https://doi.org/10.1177/0956797613498395)
- Beck, D. M. & Kastner, S. (2009). Top-down and bottom-up mechanisms in biasing competition in the human brain. *Vision Research*, 49 (10), 1154-1165. [10.1016/j.visres.2008.07.012](https://doi.org/10.1016/j.visres.2008.07.012)
- Blanke, O. (2012). Multisensory brain mechanisms of bodily self-consciousness. *Nature Reviews Neuroscience*, 13 (8), 556-571. [10.1038/nrn3292](https://doi.org/10.1038/nrn3292)
- Blanke, O., Ortigue, S., Landis, T. & Seeck, M. (2002). Stimulating illusory own-body perceptions. *Nature*, 419 (6904), 269-270. [10.1038/419269a](https://doi.org/10.1038/419269a)
- Blanke, O., Landis, T., Spinelli, L. & Seeck, M. (2004). Out-of-body experience and autoscopia of neurological origin. *Brain*, 127 (Pt 2), 243-258. [10.1093/brain/awh040](https://doi.org/10.1093/brain/awh040)
- Blanke, O. & Mohr, C. (2005). Out-of-body experience, heautoscopy, and autoscopic hallucination of neurological origin Implications for neurocognitive mechanisms of corporeal awareness and self-consciousness. *Brain Research Reviews*, 50 (1), 184-199. [10.1016/j.brainresrev.2005.05.008](https://doi.org/10.1016/j.brainresrev.2005.05.008)
- Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford, UK: Oxford University Press.
- Cheyne, J. A. & Girard, T. A. (2009). The body unbound: Vestibular-motor hallucinations and out-of-body experiences. *Cortex*, 45 (2), 201-215. [10.1016/j.cortex.2007.05.002](https://doi.org/10.1016/j.cortex.2007.05.002)
- Clark, A. (2015). Embodied prediction. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1-21). Frankfurt a. M., GER: MIND Group.

- Cullen, K. E. (2012). The vestibular system: Multimodal integration and encoding of self-motion for motor control. *Trends in Neurosciences*, 35 (3), 185-196. [10.1016/j.tins.2011.12.001](#)
- Day, B. L. & Fitzpatrick, R. C. (2005). The vestibular system. *Current Biology*, 15 (15), R583-R586. [10.1016/j.cub.2005.07.053](#)
- De Ridder, D., Van Laere, K., Dupont, P., Menovsky, T. & Van de Heyning, P. (2007). Visualizing out-of-body experience in the brain. *The New England Journal of Medicine*, 357 (18), 1829-1833. [10.1056/NEJMoa070010](#)
- Deroualle, D. & Lopez, C. (2014). Toward a vestibular contribution to social cognition. *Frontiers in Integrative Neuroscience*, 8 (16). [10.3389/fnint.2014.00016](#)
- Desimone, R. & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Reviews in Neuroscience*, 18, 193-222. [10.1146/annurev.ne.18.030195.001205](#)
- Downar, J., Crawley, A. P., Mikulis, D. J. & Davis, K. D. (2000). A multimodal cortical network for the detection of changes in the sensory environment. *Nature Neuroscience*, 3 (3), 277-283. [10.1038/72991](#)
- Ehrsson, H. H. (2007). The experimental induction of out-of-body experiences. *Science*, 317 (5841). [10.1126/science.1142175](#)
- (2012). *The new handbook of multisensory processes*. Cambridge, MA: MIT Press.
- Fernandez, C. & Goldberg, J. M. (1971). Physiology of peripheral neurons innervating semicircular canals of the squirrel monkey. II. Response to sinusoidal stimulation and dynamics of peripheral vestibular system. *Journal of Neurophysiology*, 34 (4), 661-675.
- Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 360 (2456), 815-836. [10.1098/rstb.2005.1622](#)
- Furlanetto, T., Bertone, C. & Becchio, C. (2013). The bi-located mind: New perspectives on self-localization and self-identification. *Frontiers in Human Neuroscience*, 7 (71). [10.3389/fnhum.2013.00071](#)
- Gallese, V. & Cuccio, V. (2015). The paradigmatic body. Embodied simulation, intersubjectivity and the bodily self. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Gallese, V., Fadiga, L., Fogassi, L. & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, 119 (Pt 2), 593-609. [10.1093/brain/119.2.593](#)
- Goldberg, J. M. & Fernandez, C. (1971). Physiology of peripheral neurons innervating semicircular canals of the squirrel monkey. I. Resting discharge and response to constant angular accelerations. *Journal of Neurophysiology*, 34 (4), 635-660.
- Graziano, M. S. (2013). *Consciousness and the social brain*. New York, NY: Oxford University Press.
- Graziano, M. S. & Kastner, S. (2011). Human consciousness and its relationship to social neuroscience: A novel hypothesis. *Cognitive neuroscience*, 2 (2), 98-113. [10.1080/17588928.2011.565121](#)
- Guterstam, A. & Ehrsson, H. H. (2012). Disowning one's seen real body during an out-of-body illusion. *Consciousness and Cognition*, 21 (2), 1037-1042. [10.1016/j.concog.2012.01.018](#)
- Hamilton, A. F., Brindley, R. & Frith, U. (2009). Visual perspective taking impairment in children with autistic spectrum disorder. *Cognition*, 113 (1), 37-44. [10.1016/j.cognition.2009.07.007](#)
- Hohwy, J. (2013). *The predictive mind*. Oxford, UK: Oxford University Press.
- (2015). The neural organ explains the mind. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1-23). Frankfurt a. M., GER: MIND Group.
- Husserl, E. (1950). *Cartesianische Meditationen und Pariser Vorträge*. The Hague, NLD: Martinus Nijhoff Publishers.
- Indovina, I., Maffei, V., Bosco, G., Zago, M., Macaluso, E. & Lacquaniti, F. (2005). Representation of visual gravitational motion in the human vestibular cortex. *Science*, 308 (5720), 416-419. [10.1126/science.1107961](#)
- Indovina, I., Maffei, V., Pauwels, K., Macaluso, E., Orban, G. A. & Lacquaniti, F. (2013). Simulated self-motion in a visual gravity field: Sensitivity to vertical and horizontal heading in the human brain. *NeuroImage*, 71, 114-124. [10.1016/j.neuroimage.2013.01.005](#)
- Ionta, S., Heydrich, L., Lenggenhager, B., Mouthon, M., Fornari, E., Chapuis, D., Gassert, R. & Blanke, O. (2011). Multisensory mechanisms in temporo-parietal cortex support self-location and first-person perspective. *Neuron*, 70 (2), 363-374. [10.1016/j.neuron.2011.03.009](#)
- Jeannerod, M. (2003). The mechanism of self-recognition in humans. *Behavioural Brain Research*, 142 (1-2), 1-15. [10.1016/S0166-4328\(02\)00384-4](#)
- Lenggenhager, B., Tadi, T., Metzinger, T. & Blanke, O. (2007). Video ergo sum: Manipulating bodily self-consciousness. *Science*, 317 (5841), 1096-1099. [10.1126/science.1143439](#)
- Lenggenhager, B., Mouthon, M. & Blanke, O. (2009). Spatial aspects of bodily self-consciousness. *Conscious-*

- ness and Cognition, 18 (1), 110-117.
[10.1016/j.concog.2008.11.003](https://doi.org/10.1016/j.concog.2008.11.003)
- Lenggenhager, B. & Lopez, C. (2015). Vestibular contributions to the sense of body, self and others. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Limanowski, J. & Blankenburg, F. (2013). Minimal self-models and the free energy principle. *Frontiers in Human Neuroscience*, 7 (547). [10.3389/fnhum.2013.00547](https://doi.org/10.3389/fnhum.2013.00547)
- Limanowski, J. & Hecht, H. (2011). Where do we stand on locating the self? *Psychology*, 2 (4), 312-317.
[10.4236/psych.2011.24049](https://doi.org/10.4236/psych.2011.24049)
- Lopez, C. & Blanke, O. (2011). The thalamocortical vestibular system in animals and humans. *Brain Research Reviews*, 67 (1-2), 119-146.
[10.1016/j.brainresrev.2010.12.002](https://doi.org/10.1016/j.brainresrev.2010.12.002)
- Lopez, C., Halje, P. & Blanke, O. (2008). Body ownership and embodiment: vestibular and multisensory mechanisms. *Clinical Neurophysiology*, 38 (3), 149-161.
[10.1016/j.neucli.2007.12.006](https://doi.org/10.1016/j.neucli.2007.12.006)
- Lopez, C., Bachofner, C., Mercier, M. & Blanke, O. (2009). Gravity and observer's body orientation influence the visual perception of human body postures. *Journal of Vision*, 9 (5), 11-14. [10.1167/9.5.1](https://doi.org/10.1167/9.5.1)
- Lopez, C., Blanke, O. & Mast, F. W. (2012). The human vestibular cortex revealed by coordinate-based activation likelihood estimation meta-analysis. *Neuroscience*, 212, 159-179. [10.1016/j.neuroscience.2012.03.028](https://doi.org/10.1016/j.neuroscience.2012.03.028)
- Lopez, C., Falconer, C. J. & Mast, F. W. (2013). Being moved by the self and others: Influence of empathy on self-motion perception. *PloS one*, 8 (1), e48293-e48293.
[10.1371/journal.pone.0048293](https://doi.org/10.1371/journal.pone.0048293)
- McIntyre, J., Zago, M., Berthoz, A. & Lacquaniti, F. (2001). Does the brain model Newton's laws? *Nature Neuroscience*, 4 (7), 693-694. [10.1038/89477](https://doi.org/10.1038/89477)
- Metzinger, T. (2003). *Being no one*. Boston: MIT Press.
- (2009). Why are out-of-body experiences interesting for philosophers? The theoretical relevance of OBE research. *Cortex*, 45 (2), 256-258.
[10.1016/j.cortex.2008.09.004](https://doi.org/10.1016/j.cortex.2008.09.004)
- Newen, A. & Vogeley, K. (2003). Self-representation: Searching for a neural signature of self-consciousness. *Consciousness and Cognition*, 12 (4), 529-543.
[10.1016/S1053-8100\(03\)00080-1](https://doi.org/10.1016/S1053-8100(03)00080-1)
- Petkova, V. I. & Ehrsson, H. H. (2008). If I were you: perceptual illusion of body swapping. *PLOS ONE*, 3 (8), e3832-e3832. [10.1371/journal.pone.0003832](https://doi.org/10.1371/journal.pone.0003832)
- Petkova, V. I., Bjornsdotter, M., Gentile, G., Jonsson, T., Li, T. Q. & Ehrsson, H. H. (2011). From part- to whole-body ownership in the multisensory brain. *Current Biology*, 21 (13), 1118-1122.
[10.1016/j.cub.2011.05.022](https://doi.org/10.1016/j.cub.2011.05.022)
- Petkova, V. I., Khoshnevis, M. & Ehrsson, H. H. (2011). The perspective matters! Multisensory integration in ego-centric reference frames determines full-body ownership. *Frontiers in Psychology*, 2 (35).
[10.3389/fpsyg.2011.00035](https://doi.org/10.3389/fpsyg.2011.00035)
- Pfeiffer, C., Lopez, C., Schmutz, V., Duenas, J. A., Martuzzi, R. & Blanke, O. (2013). Multisensory origin of the subjective first-person perspective: Visual, tactile, and vestibular mechanisms. *PloS one*, 8 (4).
[10.1371/journal.pone.0061751](https://doi.org/10.1371/journal.pone.0061751)
- Pfeiffer, C., Schmutz, V. & Blanke, O. (2014). Visuospatial viewpoint manipulation during full-body illusion modulates subjective first-person perspective. *Experimental Brain Research*. [10.1007/s00221-014-4080-0](https://doi.org/10.1007/s00221-014-4080-0)
- Pfeiffer, C., Serino, A. & Blanke, O. (2014). The vestibular system: A spatial reference for bodily self-consciousness. *Frontiers in Integrative Neuroscience*, 8 (31).
[10.3389/fnint.2014.00031](https://doi.org/10.3389/fnint.2014.00031)
- Rizzolatti, G., Fadiga, L., Gallese, V. & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3 (2), 131-141.
- Rizzolatti, G. & Sinigaglia, C. (2010). The functional role of the parieto-frontal mirror circuit: Interpretations and misinterpretations. *Nature Reviews Neuroscience*, 11 (4), 264-274. [10.1038/nrn2805](https://doi.org/10.1038/nrn2805)
- Sciutti, A., Demougeot, L., Berret, B., Toma, S., Sandini, G., Papaxanthis, C. & Pozzo, T. (2012). Visual gravity influences arm movement planning. *Journal of neurophysiology*, 107 (12), 3433-3445. [10.1152/jn.00420.2011](https://doi.org/10.1152/jn.00420.2011)
- Serino, A., Alsmith, A., Costantini, M., Mandrigin, A., Tajadura-Jimenez, A. & Lopez, C. (2013). Bodily ownership and self-location: Components of bodily self-consciousness. *Consciousness and Cognition*, 22 (4), 1239-1252. [10.1016/j.concog.2013.08.013](https://doi.org/10.1016/j.concog.2013.08.013)
- Seth, A. (2014). The cybernetic Bayesian brain: From interoceptive inference to sensorimotor contingencies. In T. Metzinger & J. W. Windt (Eds.) *Open MIND*. Frankfurt am Main: MIND Group.
- Stein, B. E. & Stanford, T. R. (2008). Multisensory integration: current issues from the perspective of the single neuron. *Nature Reviews Neuroscience*, 9 (4), 255-266.
[10.1038/nrn2331](https://doi.org/10.1038/nrn2331)
- Thompson, P. (1980). Margaret Thatcher: A new illusion. *Perception*, 9 (4), 483-484.
- van der Hoort, B., Guterstam, A. & Ehrsson, H. H. (2011). Being Barbie: The size of one's own body de-

- termines the perceived size of the world. *PLOS ONE*, 6 (5), e20195-e20195. [10.1371/journal.pone.0020195](https://doi.org/10.1371/journal.pone.0020195)
- van Elk, M. & Blanke, O. (2013). Imagined own-body transformations during passive self-motion. *Psychological Research*, 78 (1). [10.1007/s00426-013-0486-8](https://doi.org/10.1007/s00426-013-0486-8)
- Windt, J. M. (2010). The immersive spatiotemporal hallucination model of dreaming. *Phenomenology and the Cognitive Sciences*, 9 (2), 295-316. [10.1007/s11097-010-9163-1](https://doi.org/10.1007/s11097-010-9163-1)
- zu Eulenburg, P., Caspers, S., Roski, C. & Eickhoff, S. B. (2012). Meta-analytical definition and functional connectivity of the human vestibular cortex. *NeuroImage*, 60 (1), 162-169. [10.1016/j.neuroimage.2011.12.032](https://doi.org/10.1016/j.neuroimage.2011.12.032)

Embodied Simulation: A Paradigm for the Constitution of Self and Others

A Reply to Christian Pfeiffer

Vittorio Gallese & Valentina Cuccio

The main criticism Pfeiffer advances in his commentary is that our proposal is too narrow. Embodied simulation (ES), in his view equated to motor resonance, is not a sufficiently primary mechanism on which we can base a unified neurobiological theory of the earliest sense of self and others. According to Pfeiffer, motor resonance needs to be complemented by other more basic and primary mechanisms. Hence, as an alternative to our proposal, he suggests that multisensory spatial processing can play this role, primarily contributing to the earliest foundation of the sense of self and others. In our reply we stress on the one hand that identifying ES only with motor resonance is a partial view that may give rise to fallacious arguments, since ES also deals with emotions and sensations. We also show, on the other hand, that ES and multisensory integration should not be seen as alternative solutions to the problem of the neural bases of the bodily self, because multimodal integration carried out by the cortical motor system *is* an instantiation of ES. We conclude by stressing the role ES might have played in the transition from bodily experience to symbolic expression.

Keywords

Attention schema theory | Bodily self | Embodied simulation | Language | Motor resonance | Multimodal integration | Paradigm | Peri-personal space | Social cognition

Authors

[Vittorio Gallese](#)

vittorio.gallese@unipr.it

Università degli Studi di Parma
Parma, Italy

[Valentina Cuccio](#)

Università degli Studi di Palermo
Palermo, Italy

Commentator

[Christian Pfeiffer](#)

christian.pfeiffer@epfl.ch

Ecole Polytechnique Fédérale
Lausanne, Switzerland

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 An overview of Pfeiffer's criticisms

We would like to thank Christian Pfeiffer for his very well-articulated commentary on our paper “The paradigmatic body: Embodied Simulation, Intersubjectivity, the Bodily Self, and Language” ([Gallese & Cuccio this collection](#)). His comments and criticisms offered us the opportunity to further reflect on some of the ideas

proposed in our piece. The aim of our paper was to discuss the role of the body in the constitution of the earliest and primary sense of self and others and, also, to emphasize the constitutive role of the body in a specifically human modality of intersubjectivity: language. To be more precise, we identified a biological mech-

anism, embodied simulation (ES), as a primary source of intersubjectivity, the sense of self, and language. The mechanism of ES is widely described in the paper and its role in human cognition is explained by also resorting to the Aristotelian notion of *paradeigma*.

The commentary offered by Christian Pfeiffer is focused on a partial aspect of our much wider proposal. In fact, the author only discusses the constitutive role motor resonance has for the sense of self and for social cognition. However, motor resonance is just one dimension of the mechanism of ES. As argued in our paper and elsewhere (see [Gallese & Sinigaglia 2011a](#); [Gallese 2014](#)) the mechanism of simulation is widespread in the brain and it also characterizes the nervous structures involved in the experience of emotions and sensations. All these dimensions of ES should be taken into account. To identify ES only with motor resonance is a partial view that may give rise to fallacious arguments.

The main criticism Pfeiffer advances in his commentary is that our proposal for the constitutive role of motor resonance is too narrow. ES, in his view equated to motor resonance, cannot be the primary neurobiological mechanism at the basis of both the sense of self and others. According to Pfeiffer, motor resonance needs to be complemented by other more basic and primary mechanisms. Hence, as an alternative to our proposal, he suggests that multisensory spatial processing can play this role, primarily contributing to the earliest foundation of the sense of self and others. To support this claim, he provides theoretical arguments and presents empirical data structured in three different sections. Each of these sections supposedly provides evidence of the role of multisensory spatial processing in the foundation of a bodily sense of self and others.

In the first section Pfeiffer addresses the issue of intersubjectivity and presents the Attention schema theory (AS). In his proposal, our ability to understand others is primarily based on a mechanism more primitive than ES-as-motor-resonance: spatial coding of attention. AS predicts that we understand the current state of awareness of our conspecifics by means

of schematic representations of their states of attention ([Pfeiffer this collection](#), p. 4). In other words, according to AS, by using a representation of the spatial relationship between the individual we are observing and the spatial focus of her/his attention we can likely predict his intentions and, as a consequence, his actions. [Pfeiffer \(this collection](#), p. 4) also discusses recent empirical findings on the neural structures underlying the AS. It seems that the neural structures for the spatial coding of attention are based in the right temporo-parietal junction (TPJ) and in the superior temporal sulcus (STS). These neural structures do not overlap with the neural circuits involved in ES.

In the second section Pfeiffer addresses the issue of the bodily foundation of the sense of self. The experience of being a bodily self can be decomposed into four different aspects ([Pfeiffer this collection](#), p. 5): body ownership, self-location, first-person perspective, and agency. According to Pfeiffer, motor resonance can account only for body ownership and agency, directly contributing to these (non-spatial) aspects of the bodily self. However, for the two spatial components of the bodily self we need a different account. In fact, according to Pfeiffer, empirical evidence suggests that these spatial aspects of the bodily self, which imply multisensory spatial representations, are encoded in a brain region, the TPJ, not characterized by motor resonance. Hence, motor resonance, while being still necessary for the bodily foundation of some basic aspects of the self, is not a sufficiently primary mechanism, since different neural structures are also needed for the bodily foundation of the self. In support of this claim, Pfeiffer discusses data from neurological patients with out-of-body experiences and other kinds of altered states.

Finally, in the third section the constitutive role of the vestibular system to the bodily foundation of both the consciousness of self and others is discussed. It is proposed that this system, which encodes gravity and head motion and is associated with multisensory spatial processing, significantly and primarily contributes to our ability to distinguish between motions of our own body and motions of other

people's bodies, in this way contributing to both the foundation of the sense of self and social cognition. Empirical studies are reported to support these claims. In addition, empirical data showing that the vestibular cortical network overlaps with neural structures underlying the bodily foundation of both the sense of self and others, as discussed in the two previous sections, are presented.

In the light of the empirical evidence discussed in his commentary, Christian Pfeiffer concludes that ES-as-motor-resonance is not a sufficiently primary mechanism on which we can base a unified neurobiological theory of the earliest sense of self and others. In the next section we answer these criticisms.

2 Responses

First, we would like to point out that ES is not confined to motor resonance of others' actions, like that instantiated by macaques' mirror neurons, as in humans ES also encompasses the activation of somatosensory areas during the observation of others' tactile experiences, the activation of pain-related areas like the anterior insula and the anterior cingulate cortex during the observation of others' pain, and the activation of the anterior insula and limbic structures like the amygdala during the observation of others' emotions like disgust and fear (see our paper, p. 9 and Gallese & Sinigaglia 2011a). Thus, motor resonance only describes one partial aspect of ES.

Two distinct arguments can be used to explain why we do not think that AS constitutes a valid alternative to ES, as argued by Pfeiffer. We certainly agree with Pfeiffer that shared attention, that is, the capacity to direct the gaze to an object gazed by someone else, is a basic ingredient of social cognition. Indeed, as maintained by Colwyn Trevarthen (1977), shared attention marks in human infants around 9 months of age the transition from primary to secondary intersubjectivity. However, shared attention constitutes only one aspect of intersubjectivity and social cognition, thus AS at best only covers a partial aspect of social cognition and therefore appears to be more limited than

ES in this respect. Furthermore, and most importantly, shared attention can be linked to motor resonance. Shepherd, Klein, Deaner, and Platt (2009) discovered in macaques a class of mirror neurons in the lateral intraparietal (LIP) area involved in oculomotor control, signaling both when the monkey looked at a given direction in space and when it observed another monkey looking in the same direction. These authors suggested that LIP mirror neurons for gaze might contribute to the sharing of observed attention. This evidence shows that shared attention is not divorced from motor resonance, but actually requires it.

A further argument in our opinion demonstrates that ES and AS should not be seen as alternative solutions to the problem of social cognition. Multisensory integration is a pervasive feature of parieto-frontal centers involved in sensory-motor planning and control. Indeed an influential theory about attention, the "Premotor Theory of Attention" (see Rizzolatti et al. 1987; Rizzolatti et al. 1994) states that spatial attention results from the activation of the same "pragmatic" circuits that program oculomotor behavior and other motor activities, even if such activation does not produce any overt motor behavior, thus qualifying as motor simulation.

We would like to emphasize even more strongly than we did in the paper that a crucial role of the cortical motor system is precisely that of integrating multiple sources of body-related sensory signals, like tactile, visual and auditory stimuli (see our paper, pp. 10–11; see also Gallese & Sinigaglia 2010, 2011b; Gallese 2014). The ventral premotor cortex (vPMC) might represent one of the essential anatomofunctional bases for the motor aspect of bodily selfhood, specifically because of its role in integrating self-related multisensory information. This hypothesis is corroborated by clinical and functional evidence showing the systematic involvement of vPMC with body awareness (Ehrsson et al. 2004; Berti et al. 2005; Arzy et al. 2006). This evidence demonstrates a tight relationship between the bodily self-related multimodal integration carried out by the cortical motor areas specifying the motor potentialities of one's body and guiding its motor behavior

and the implicit awareness one entertains of one's body as one's own body and of one's behavior as one's own behavior.

The vPMC is anatomically connected to visual and somatosensory areas in the posterior parietal cortex and to frontal motor areas and for this reason it is plausible to assume that vPMC activity reflects the detection of congruent multisensory signals related to one's own body parts: this mechanism could be responsible for the feeling of body ownership. The motor aspects of the bodily self-enable the integration of self-related multimodal sensory information about the body and about the world with which the body interacts, as epitomized by the properties of macaques' premotor neurons in area F4 (see [Fogassi et al. 1996](#); [Rizzolatti et al. 1997](#)) and the analogous functional properties displayed by the human homologue of area F4 (see [Bremmer et al. 2001](#)). The same neurons controlling the movement in space of the head or of the upper limb also respond to tactile, visual, and auditory stimuli, provided they are applied to the same body part, like tactile stimuli, or they occur in the body-part-centered peri-personal space, like visual and auditory stimuli. Thus, we think that ES and multisensory integration should not be seen as alternative solutions to the problem of the neural bases of the bodily self, because multimodal integration carried out by vPMC *is* an instantiation of ES. We agree with Pfeiffer, however, that other brain areas, like TPJ, might contribute to a coherent sense of one's own body. It must be added that TPJ is part of a network (including the posterior parietal cortex, and the premotor cortex) implicated in multisensory integration during self-related and other-related events and experiences. Indeed, as shown by [Ebisch et al. \(2011\)](#), the observation of others' affective tactile experiences leads to the activation of observers' vPMC and second somatosensory area and to the inactivation of observers' posterior insula. Functional connectivity revealed a significant interaction between the posterior insula, right TPJ, left pre-central gyrus, and right posterior parietal cortex during the observation of other's affective touch. These data suggest that TPJ might be involved in mapping the self-other dif-

ferentiation, by means of lower-level computational mechanisms for generating, testing, and correcting internal predictions about external sensory events.

Last, we agree with Pfeiffer that the vestibular system might contribute to the bodily foundation of both the consciousness of self and others and we thank him for having pointed this out, thus integrating our perspective.

3 Conclusions

It seems that the data discussed in the previous section allow us to come to the conclusion that ES is the primary and earliest mechanism contributing to the foundation of the sense of self and others. That said, in conclusion, we would like to stress again the issue of the cognitive role ES has in relation to language. Though the aspect of the relation between ES and language was not addressed in Pfeiffer's commentary, this was a central point of our proposal. The relation between ES and language is two-sided. On the one hand, empirical evidence has shown the role ES plays in language comprehension. These data (for an overview see [Gallese & Cuccio this collection](#), p. 13) suggest that the bodily, sensory, and motor dimensions play a constitutive role in language, both ontogenetically and phylogenetically. On the other hand, being linguistic creatures, we humans are the only living species able to fix and relive specific aspects of our bodily experiences by means of symbols. Words or other forms of symbolic representations such as art, for example, allow us to activate and relive our bodily experiences. In this way, by means of symbolic representations, we can share our bodily experiences, enacted by ES, even with people far away from us in time and space. As argued in our paper, ES is a model of our own experiences and its defining features are best explained by resorting to the Aristotelian notion of *paradeigma*. ES-as-*paradeigma* (and not just as motor resonance) provides a neurobiologically-based new perspective on human social cognition and ultimately on the very definition of human nature.

Acknowledgments

This work was supported by the EU Grant Towards an Embodied Science of InterSubjectivity (TESIS, FP7-PEOPLE-2010-ITN, 264828) and by the KOSMOS Fellowship from Humboldt University, Berlin to VG.

References

- Arzy, S., Overney, L. S., Landis, T. & Blanke, O. (2006). Neural mechanisms of embodiment: Asomatognosia due to premotor cortex damage. *Archives of Neurology*, 63 (7), 1022-1025. [Neural mechanisms of embodiment: Asomatognosia due to premotor cortex damage](#)
- Berti, A., Bottini, G., Gandola, M., Pia, L., Smania, N., Stracciari, A., Castiglioni, I., Vallar, G. & Paulesu, E. (2005). Shared cortical anatomy for motor awareness and motor control. *Science*, 309 (5733), 488-491. [10.1126/science.1110625](#)
- Bremmer, F., Schlack, A., Shah, N. J., Zafiris, O., Kubischik, M., Hoffmann, K., Zilles, K. & Fink, G. R. (2001). Polymodal motion processing in posterior parietal and premotor cortex: a human fMRI study strongly implies equivalencies between humans and monkeys. *Neuron*, 29 (1), 287-296. [10.1016/S0896-6273\(01\)00198-2](#)
- Ebisch, S. J. H., Ferri, F., Salone, A., d'Amico, L., Perucci, M. G., Ferro, F. M., Romani, G. L. & Gallese, V. (2011). Differential involvement of somatosensory and interoceptive cortices during the observation of affective touch. *Journal of Cognitive Neurosciences*, 23 (7), 1808-1822. [10.1162/jocn.2010.21551](#)
- Ehrsson, H. H., Spence, C. & Passingham, R. E. (2004). That's my hand! Activity in premotor cortex reflects feeling of ownership of a limb. *Science*, 305 (5685), 875-877. [10.1126/science.1097011](#)
- Fogassi, L., Gallese, V., Fadiga, L., Luppino, G., Matelli, M. & Rizzolatti, G. (1996). Coding of peripersonal space in inferior premotor cortex (area F4). *Journal of Neurophysiology*, 76 (1), 141-157.
- Gallese, V. (2014). Bodily Selves in Relation: Embodied simulation as second-person perspective on intersubjectivity. *Philosophical Transactions of the Royal Society, London. Series B Biological Sciences*, 369 (1644), 20130177-20130177. [10.1098/rstb.2013.0177](#)
- Gallese, V. & Cuccio, V. (2015). The Paradigmatic Body. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1-23). Frankfurt a. M., GER: MIND Group.
- Gallese, V. & Sinigaglia, C. (2010). The bodily self as power for action. *Neuropsychologia*, 48 (3), 746-755. [10.1016/j.neuropsychologia.2009.09.038](#)
- (2011a). What is so special with Embodied Simulation. *Trends in Cognitive Sciences*, 15 (11), 512-519. [10.1016/j.tics.2011.09.003](#)
- (2011b). How the body in action shapes the self. *Journal of Consciousness Studies*, 18 (7-8), 117-143.
- Pfeiffer, C. (2015). Multisensory Spatial Mechanisms of the Bodily Self and Social Cognition. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1-14). Frankfurt a. M., GER: MIND Group.
- Rizzolatti, G., Riggio, L., Dascola, I. & Umiltà, C. (1987). Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention. *Neuropsychologia*, 25 (1a), 31-40.
- Rizzolatti, G., Riggio, L. & Sheliga, B. M. (1994). Space and selective attention. In C. Umiltà & M. Moscovitch (Eds.) *Attention and performance XV* (pp. 231-265). Cambridge, MA: MIT Press.
- Rizzolatti, G., Fadiga, L., Fogassi, L. & Gallese, V. (1997). The space around us. *Science*, 277 (5323), 190-191. [10.1126/science.277.5323.190](#)
- Shepherd, S. V., Klein, J. T., Deaner, R. O. & Platt, M. L. (2009). Mirroring of attention by neurons in macaque parietal cortex. *Proceedings of The National Academy of Sciences USA*, 106 (23), 9489-9494. [10.1073/pnas.0900419106](#)
- Trevarthen, C. (1977). Descriptive analyses of infant communicative behavior. In H. R. Schaffer (Ed.) *Studies in mother-infant interaction* (pp. 227-270). London, UK: Academic Press.

All the Self We Need

Philip Gerrans

I use evidence from psychiatric disorders involving the experience of depersonalisation to decompose the causal and cognitive structure of experiences reported as self-awareness. I combine insights from predictive coding theory and the appraisal theory of emotion to explain the association between hypoactivity in the Anterior Insular Cortex and depersonalization. This resolves a puzzle for some theories raised by the fact that reduced affective response in depersonalization is associated with normal interoception and activity in Posterior Insular Cortex. It also elegantly accounts for the role of anxiety in depersonalisation via the role of attention in predictive coding theories.

Keywords

Affective processing | Appraisal theory of emotion | Bodily awareness | Depersonalisation | Disorders of self-awareness | Identity | Phenomenal avatar | Predictive coding | Self | Simulation

Author

[Philip Gerrans](#)
philip.gerrans@adelaide.edu.au
University of Adelaide
Adelaide, Australia

Commentator

[Ying-Tung Lin](#)
lingyingtung@gmail.com
國立陽明大學
National Yang-Ming University
Taipei, Taiwan

Editors

[Thomas Metzinger](#)
metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

“Who is the I that knows the bodily me, who has an image of myself and a sense of identity over time, who knows that I have appropriate strivings?” I know all these things, and what is more, I know that I know them. But who is it who has this perspectival grasp? It is much easier to *feel* the self than to *define* the self ([Allport 1961](#), p. 128)

1 Preliminary remarks

I think Allport has it the wrong way round. It is easy to *define* the self, as he in fact does, as the thing that thinks, feels, perceives and has a sense of identity over time. It is hard, however, to find an entity that fits the definition. This is so even though, according to Allport, experiencing being a self is unproblematic (“it is easier to *feel* the

self”). In fact, the experience of being someone is actually very elusive, phenomenologically and conceptually. On some accounts self-awareness is actually the experience of *Being No-One*¹ ([Met-](#)

¹ Strictly speaking, the experience is not of being no one, since there is no one to be. Rather it is an experience we cannot help but take to be of being someone, even though there is no entity causing the experience.

zinger 2003). In this chapter I use disorders of self-awareness to develop an account of the experience which gives rise to the feelings referred to by Allport. In the final sections we shall see whether our experience is of being someone, no-one, or something other than a self. Perhaps a body. Or the process of thinking.

The conclusion is that self-awareness is *almost* a necessary or inevitable illusion when the mind is functioning smoothly. The experience of being a self is produced by mechanisms that compute the relevance of sensory (including, and especially, bodily) information to a variety of organismic goals represented at different levels of explicitness in a cognitive hierarchy. The computations relate information to those goals, *not to selves*. Those computations of goal relevance produce consequent bodily feelings. Those, and only those, feelings give us the phenomenal information we need to plan, remember, and interact with other people and the world as though we are unified selves. Thomas Metzinger argues that integration of information in experience amounts to the construction of a phenomenal avatar, which the brain uses to manoeuvre the organism through the world (Blanke & Metzinger 2009; Metzinger 2011). I agree, and the rest of the chapter can be seen as an attempt to anatomise that avatar. I use evidence from psychiatric disorders involving the experience of depersonalisation to decompose the causal and cognitive structure of experiences reported as self-awareness.

2 Introduction

So many psychiatric disorders are explained in terms of the way the patient experiences herself that, even if intuitive or philosophical theories which posit a self as the object of experience are not correct, there is an interesting phenomenon there to be explained. My idea is that the best integrative explanation of those disorders is *ipso facto* the best philosophical theory of self-awareness because those disorders cannot be explained other than via a model of the way the

experience is generated in normal and abnormal situations.² Once we have explained those disorders we can determine the theoretical utility of overlapping folk, clinical and philosophical conceptions of self-awareness. Thus, the approach I take is consistent with that proposed by Dominic Murphy in his plea for a (cognitive neuro) scientific psychiatry: “we arrive at a comprehensive set of positive facts about how the mind works, and then ask which of its products and breakdowns matter for our various projects” (2006, p. 105).

So until the concluding sections I use the term self-awareness to refer to the experience we report in terms of awareness of being a unified persisting entity: the same person at a time and over time. It may turn out that such experiences are illusions or misinterpretations of some other phenomenon, perhaps because there are no such entities as selves, but I delay that discussion until the evidence is assembled. To anticipate, I think the intuitive folk concept of self-awareness is very like the intuitive concept of episodic memory, which is of “re-experiencing” a previous episode. Cognitive neuroscience tells us that in fact episodic memory experiences are constructed to suit current cognitive context rather than retrieved intact. However it does no harm in everyday life to think of episodic memory as content-preserving retrieval of past experience. Similarly the intuitive conception of self-awareness tracks processes which, when they function harmoniously, produce experiences that provide a plausible basis for the concept of a unified and persisting self. That

2 In other words I take the strong view advocated by Murphy. The ontology of the mind *is* the ontology of cognitive science. The reason is that only with the correct theory of cognitive architecture in place can we understand how neural processes implement the cognitive processes whose operations we experience as personal-level phenomenology. That personal-level phenomenology provides the raw material for intuitive or folk explanations that abstract from cognitive and neural realization. But that abstraction is precisely why, as Halligan and Marshall once memorably said, in the absence of a suitably constrained cognitive model, psychiatry will be consumed by “the expensive and extensive search for non-existent entities” (Halligan & Marshall 1996, p. 6). I take the view that mechanistic (in the sense of neuroscientific) and phenomenological (based on reflection on the nature of experience) explanation are not independent projects. One *could* have a purely personal-level phenomenological ontology of mind. But the fact that such ontologies mislead about the sources of psychiatric disorder is a reason to search for an integrative theory. But the only way neuroscience can explain experience is via a detailed computational, cognitive theory.

There is no substantial Cartesian, or bodily, or neural, entity that sustains the properties ascribed by Allport. Thus part of Metzinger’s project is to explain why we feel as though we are substantial entities.

concept, while not entirely accurate, provides a useful ability to represent and communicate sufficient unity and persistence. If I tell you I will be happy to pick you up at the airport you need to be able to rely on *me* to be at the Arrivals gate. The precise nature of my (dis)unification as a single self is not relevant. If I told you I would send my body but would not be present myself you would phone a psychiatrist. (It would be super to be able to deputise your body to attend departmental meetings, weddings etc. on your behalf, wouldn't it?) Yet something like that phenomenon of alienation occurs in depersonalisation, as a deeply felt and distressing phenomenon. The difference in experience between people with depersonalisation and those without it is an essential *explanandum* both for psychiatry and for philosophers interested in the (possibly illusory) phenomenology of selfhood.

The rest of the chapter proceeds as follows. I first discuss the Cotard delusion, in which people say that they have died, disappeared or do not exist (*déire de négation*). The Cotard delusion raises a set of questions about the relationship between self-awareness, bodily experience, and affective processing. I outline some suggestive intuitive answers to these questions based on the phenomenology of the disorder but argue that they are insufficient as explanations. A deeper explanation is provided by the cognitive neuroscience of depersonalisation. That explanation relies on a theoretical framework that draws on

- I. The appraisal theory of emotion
- II. The simulation model of memory and prospection
- III. The hierarchical predictive coding model of cognitive processing

This framework allows us to explain how:

- affective experiences provide the basis for self-awareness as a distinct form of bodily awareness *moment to moment*
- those moment to moment experiences of self-awareness can be annexed to cognitive processes whose temporal reach is longer than

the present, creating the experience/illusion of a continuing self

- when affective processing is compromised the resultant experience is reported as change, or in extreme cases, loss, of self. Mere absence of bodily or affective response *per se* does not lead to depersonalisation. What leads to depersonalisation is the absence of *predicted* affective responses that normally constitute self-awareness that leads to depersonalisation. This explanation also provides a full explanation of an intriguing phenomenological observation made by Cotard about the role of anxiety in generating depersonalisation.

With this theoretical framework in place I discuss depersonalisation disorder and depersonalisation aspects of the Cotard delusion, resolving some of the questions raised by the initial phenomenological explanation.

Once those questions are answered we can make some comments on the theoretical utility of philosophical theories of self-awareness, which for convenience I classify into four types: Illusory Self, Fat Controller, Embodied Self, Narrative Self. The Illusory Self is a version of the Humean idea that self-awareness is either illusory or a theoretically loaded misdescription of some other experiential phenomenon (perceptual, interoceptive, emotional, somatic). It is quite consistent with the Illusory Self theory that the experience is a “necessary illusion” created by architecture installed by evolution. The Fat Controller theory is that self-awareness is the experience of a genuine *substantial* self, a locus of higher order cognitive integration and top down control (like the aptly-named Will Self's Fat Controller in his *Quantity Theory of Insanity*). Embodied Self theories identify self-awareness with forms of bodily awareness. Finally there are Narrative views of the self, thin and thick. On the thin view the self is a “centre of narrative gravity”, a fictive entity generated by the Joycean machine to organize and communicate. On thicker views the self is not a fiction but a genuine cognitive entity whose essence is to construct and communicate its own autobiography as an essential aspect of higher order cognitive control. The Thin view goes

naturally with the Illusory Self view: it explains the persistence of the Illusion, while the Thick view (naturally enough) fits well with Fat Controller views.

Cognitive neuroscience does not vindicate any of these theories. However this does not mean that we should regard the phenomenon of self-awareness as empirically disconfirmed. It turns out that there are cognitive processes that generate experiences with some of the properties ascribed by different theories under different conditions. So, as with episodic memory, rather than explaining self-awareness away, we can describe and explain the nature of the experiences reported as self-awareness in terms of the structure of the processing which generates it. Self-awareness is a cognitive illusion, based on the nature of affective processing. The relevant experience plays a crucial role in higher levels of cognitive control that organise and communicate experience in narrative form: fragments, episodes, chronicles, histories and epics (Currie & Jureidini 2004; Goldie 2011; Jureidini 2012). This conjunction of processes makes self-awareness an irresistible illusion. The nature and necessity of this illusion is shown by the nature of the disorders that arise when it fails.

3 The phenomenology of the Cotard syndrome

In their study of uncommon psychiatric syndromes Enoch & Trethowan (1991) provided a haunting clinical vignette. They described a patient who said that her body was decomposing and disappearing and that eventually she would be “just a voice”. Another patient suffering from the same condition described himself as a “dead star” orbiting an inert galaxy. The Cotard delusion, from which these patients suffer, was described by Jules Cotard in 1882 as a “*délire de négation*”, a delusion of inexistence (Cotard 1880, 1882, 1884, 1891; Debruyne et al. 2009). It is also described as a paradoxical belief that one is dead. The current cultural fascination with zombies provides the metaphor of “walking corpse” syndrome to describe the condition. However, as with many psychiatric disorders, perhaps the most telling descriptions and ex-

planations of the phenomenon were provided in the nineteenth century, in this case by Cotard himself. He described his patient thus:

Miss X affirms she has no brain, no nerves, no chest, no stomach, no intestines; there’s only skin and bones of a decomposing body. . . . She has no soul, God does not exist, neither the devil. She’s nothing more than a decomposing body, and has no need to eat for living, she cannot die a natural death, she exists eternally if she’s not burned, the fire will be the only solution for her. (Translation from Cotard 1880)

Cotard explained this delusion as a consequence of a particular type of psychotic depression “characterized by anxious melancholia, ideas of damnation or rejection, insensitivity to pain, delusions of nonexistence concerning one’s own body, and delusions of immortality” (Debruyne et al. 2009, p. 67).

More recently (Gerrans 2000, 2001; Debruyne et al. 2009) the delusion of inexistence has been explained as a consequence of the experience of depersonalisation. The delusion is a personal level response to an intractable and impenetrable loss of affective response to the world. Of course to say that an experience is of depersonalization is not an explanation but an intuitive characterization: the concept expresses the phenomenology of feeling disconnected from the world including one’s own body, as though experiences are “not happening to me”. Such feelings plausibly originate in what we might call affective derealisation: the failure of emotionally salient events to trigger affective responses in the patient so that the world feels strange and unreal. Since affective responses are a form of bodily experience it makes sense that the Cotard delusion is often expressed as beliefs about alteration in body state: in particular that the body is vanishing, disappearing or dead. And since there is an intimate connection between felt body state and self-awareness this loss of normal affective response is expressed as the idea that the self no longer exists.

But surely it is equally intuitively plausible that a person suffering from derealisation might express the experience by saying that the world (perhaps including her body) feels strange, emotionally inert or unreal? In other words, why does the patient not report derealisation, the feeling that the world is unreal? One possible answer is contained in the following suggestion:

Cases of the Cotard delusion have been reported . . . in which the subject proceeds beyond reporting her rotting flesh or her death to the stage of describing the world as an inert cosmos whose processes she merely registers without using the first-person pronoun....The patient does not recognize experiences as significant for her because, due to the global suppression of affect [ex hypothesi a consequence of extreme depression], she has no qualitative responses to the acquisition of even the most significant information. These extreme cases of the Cotard delusion are those in which neural systems on which affect depends are suppressed and, as a consequence, it seems to the patient as if her experiences do not belong to her. Thus the patient reports, not changes in herself, but changes in the states of the universe, one component of which is her body, now thought of as another inert physical substance first decomposing and finally disappearing. (Gerrans 2000)

My earlier self suggested that when the patient experiences global affective suppression she experiences her body as simply a body, a physical substance rather than the body which sustains the self or the body *qua* self: Hence the depersonalisation. However this simply begs the question. What is it about affective processing which transforms representations of body states to representations of states of a self?

4 Feelings of self-relevance

Appraisal theory is familiar to theorists of emotion as the theory that emotions are representa-

tions of the significance of events for the organism. Fear, for example, results from the representation of objects as dangerous for the organism. Early appraisal theorists assimilated these appraisals to beliefs about the properties of the objects of emotion (Kenny 1963; Solomon 1976, 1993). Consequently appraisal theory has been criticized as overly intellectualistic and as ignoring the felt aspect of emotion. Fear is a visceral state whose essence is a feeling, not a judgment, runs the objection. Equally an emotional feeling may arise or persist in the absence of, or in opposition to, a judgment.

Recent versions of the theory avoid this objection by recognising that most emotional appraisals are in fact conducted by neural circuits that automatically link perception to the automatic regulation of visceral and bodily responses. Consequently appraisals issue almost instantaneously in feelings that reflect the nature of that appraisal. When we recognize a familiar person and see her smile, for example, the significance of that information for us has been represented and that representation used to initiate our own bodily response within a few hundred milliseconds (Adolphs et al. 2002; Sander et al. 2003; Sander et al. 2005; N'Diaye et al. 2009; Adolphs 2010).

The consequence of these appraisals is autonomically-regulated body states and action tendencies that produce changes in visceral and bodily state. These changes are sensed as affective feelings via specialised circuitry that evolved to monitor organismic state. At any given moment we experience a “core affect” which is the product of multiple appraisals along different dimensions at different time scales.

These affective processes essentially represent the significance of incoming information for the organism along a number of different dimensions—hedonic, prudential, dangerous, noxious, nourishing, interesting, and so on. These representations, however, relate an aspect of organismic functioning to a represented object; they do not represent a self *per se*. The detection of danger alerts the organism to the need for avoidance, for example. The consequent feeling of fear is a way of sensing the bodily consequences of that appraisal. The self as an en-

tity need not be represented in either the initial appraisal or the consequent experience. The self-relevance (as appraisal theorists call it) of dangerous objects is however *implicitly represented* in the bodily experience of fear. The same is true of all affective experiences: they carry important information about the world and the way the organism is faring in it in virtue of the appraisal processes which generate them. But they do so without representing a self in any substantial sense. Rather they relate salient information to organismic goals represented at different levels of explicitness for different purposes (Tomkins 1962, 1991; Scherer 2004).

Cognitive neuroscience has identified circuits that function as “hubs” of distributed circuits that determine the subjective relevance of information. Lower-level hubs, of which the amygdala is a central component, implement rapid online appraisals (Sander et al. 2003; Adolphs 2010) and coordinate visceral and bodily responses. These lower level hubs associate affective experiences with online sensorimotor processing of the type often described as reflexive: that is initiated by, and dependent on, encounters with the environment. It follows that such experiences decay with the representation of the stimulus. They are stimulus dependent. Such reflexive affective processes can of course only sustain a feeling of self-relevance moment to moment.

5 Simulation, affective sampling, and the self

By self-awareness, however, philosophers have in mind the experience of being an entity that exists through time, which is not something that can be produced by reflexive processing. The organism needs to be able to represent itself, not just moment-to-moment but as an entity with a history and a future (“to consider itself the same thing at a time and over time”). It must therefore be able to link affective experience to memory and prospection in the same way as it links it to perception and sensory processing moment to moment. That is to say that it must be able to appraise episodes of memory and foresight for self-relevance.

Because the temporal window of human cognition extends beyond the present we have evolved systems that recapitulate important aspects of reflexive affective processing for those higher level cognitive processes involved in planning, recollection, prospection, and decision-making. These systems *simulate* temporally distant experiences by rehearsing some of the same perceptual and emotional mechanisms activated by the simulated situation. As a result we can recall previous episodes of experience and imagine future episodes of experience and link those simulations to other high-level cognitive processes in order to plan and decide. We remember being sunburnt and imagine getting skin cancer when deciding whether to go to the beach at noon (Gusnard et al. 2001; Buckner et al. 2008; Fair et al. 2008; Broyd et al. 2009).

These simulations are the raw material of autobiographical narratives whose structure and duration can vary depending on cognitive context. They may be as simple as recall of a single event that triggers a flash of affect, but can also be assembled into elaborate histories and imaginative rehearsals depending on the cognitive context. This narrative capacity provides a crucial aspect of cognitive control possibly unique to humans. The most important aspect of these simulations is sometimes overlooked in studies that emphasise their quasi-perceptual content. That is the fact that the simulation of perceptual and sensory experience evokes affective associations. We simulate a scene in order to evoke the affective responses that represent the significance of events and objects for us. When we imagine or recall an episode of experience its affective significance is also represented in experience via the offline rehearsal of affective processing. The ventromedial prefrontal cortex is a structure which “traffics” or makes available the affective information. In effect, the ventromedial prefrontal cortex recapitulates at a higher level the properties of the amygdala. In so doing it associates affective information with explicitly represented information used in reflective decision making and planning (Ochsner et al. 2002; Bechara & Damasio 2005). It thus allows the subject to make explicit reflective appraisals. When I lie on the beach I have pleas-

ant feelings produced by low-level appraisal systems. When I imagine or recall lying on the beach while trying to decide whether to holiday in Thailand or Senegal my ventromedial prefrontal cortex makes available the affective information prompted by that simulation.

This is why “pure” episodic memory studies (such as recall of content of visual scenes) do not activate the ventromedial prefrontal cortex, whereas “activations in the ventromedial PFC [prefrontal cortex] ... are almost invariably found in *autobiographical* memory studies” (Gilboa 2004, p. 1336; my emphasis). Gilboa (2004) suggests that this is because “autobiographical memory relies on a quick intuitive ‘feeling of rightness’ to monitor the veracity and cohesiveness of retrieved memories in relation to an activated self-schema.” This is consistent with studies showing activity in the ventromedial and related subcortical structures when people make intuitive (that is, rapid and semiautomatic) judgments about themselves. When people make judgments about themselves using semantic knowledge and symbolic reasoning, ventromedial structures are less active.

This idea is supported by studies of patients with lesions to the ventromedial prefrontal cortex. These patients oscillate between various forms of reflexive cognition and more abstract forms of thinking using semantic knowledge and procedural reasoning. What they have lost is the ability, provided by ventromedial structures, to simulate affective and motivational response in the absence of the stimulus, while they retain the ability to process information in an abstract way. Consequently, a ventromedial patient may be able to do a utility calculation about her personal future but be unable to act on that knowledge. It appears that semantic knowledge is motivationally inert. Such results are often used to emphasize the necessary role of affect in deliberation, but they also suggest that what those affective responses do is provide the necessary *personal* perspective on information. They make the information *mine*, so to speak. Furthermore, this diminishment is not just *at* a time, but over time. These patients, although not amnesic in the strict sense of the term, have very limited ability for

autobiographical recall or prospection. They have no sense of a persisting self (Damasio 1994; Bechara & Damasio 2005; Gerrans & Kennett 2010).

This suggests that disorders in which people feel a diminished sense of self would be characterized by hypoactivity in the ventromedial prefrontal cortex. In a review of the neuropsychological and imaging literature, Koenigs & Grafman concluded that “one could conceive of the VMPFC patients’ selective reduction in depressive symptoms as a secondary effect of a *primary lack of self-awareness and self-reflection*” (2009, p. 242; my emphasis). In other words, patients with ventromedial damage do not “feel” personally affected when considering even quite distressing events because they cannot access or activate the required affective responses.

It seems that “mine-ness” of experience is a cognitive achievement mediated by the ventromedial prefrontal cortex. As we noted above the ventromedial prefrontal cortex is suited to play this role because it recapitulates at a higher level many of the processing properties of lower-level hubs of emotional processing that represent self-relevance. Rather than reinvent the cognitive wheel for controlled processing, evolution has provided pathways that traffic affective and reward-predictive information processed automatically at lower levels to controlled processing coordinated by the ventromedial prefrontal cortex.

In effect, these studies suggest that in both online reflexive and offline reflective processing affective processes are needed to represent the significance of the information for the subject, and it is the consequent bodily feelings that produce the feeling of self-awareness. My version of this view is in some ways an amalgam of ideas found in Seth (2013) and Proust (2013). All three of us share the view that the mind is hierarchically organized, and that feelings of self-awareness emerge when higher order, metacognitive processes such as planning or deliberation integrate bodily information which signals relevance. On Seth’s and my view the Anterior Insular Cortex (AIC) is in some ways specialized for that function in view of its architecture: it does

not merely relay first order bodily information but is involved in the representation of the significance of that information. Thus it is well placed to be the source of some of the metacognitive feelings identified by Proust (2013) as serving crucial indicator functions.³

Affective processes represent the relevance of information for an organism and initiate suitable action tendencies and autonomic responses. The bodily consequences are sensed and summarised by specialised systems that inform the organism how it is faring in the world: this is affective information (Prinz 2004). This affective information is made available to other cognitive processes, which operate at different time scales, from instantaneous and automatic, to reflective and controlled. We are able to think and behave as continuing entities because the salience of information for different organismic goals is represented by affective processes at different time scales and levels of explicitness. An organism that can *use that affective information in the process* is a *self*.

This suggests that if the ability to access affective information is lost then self-awareness would also be diminished. Thus as we suggested above a key to the experience of depersonalisation in the Cotard delusion is the profound loss of affect associated with extreme depression. This suggestion is almost correct but it ignores another stage in the production of depersonalisation. After all, from what we have said so far affective processes represent the self-relevance of information. If the consequent feelings are unavailable the world should feel not significant for the subject. That is to say the subject might feel detached from the world or as if the world was emotionally inert. But it seems an extra step from a lack of affective experience to the feeling or thought of non-existence. Of course the step might be a small one. This was the

idea of Gerrans in his pioneering work at the dawn of the millennium. He suggested that there was such an intimate connection between affective experience and the self that any profound involuntary change in affect would be felt as a change to the self. However since then interesting work on depersonalisation disorder has provided a deeper understanding of the phenomenon. That work draws on the predictive coding theory of cognitive function.

6 The predictive coding hierarchy

The mind is organized as a hierarchical system that uses representations of the world and its own states to control behavior. According to recently influential Bayesian theories of the mind, all levels of the cognitive hierarchy exploit the same principle: error correction (Friston 2003; Hohwy et al. 2008; Jones & Love 2011; Clark 2012, 2013; Hohwy 2013). Each cognitive system uses models of its domain to *predict* its future informational states, given actions performed by the organism. When those predictions are satisfied, the model is reinforced; when they are not, the model is revised or updated, and new predictions are generated to govern the process of error correction. Discrepancy between actual and predicted information state is called *surprisal* and represented in the form of an error signal. That signal is referred to a higher-level supervisory system, which has access to a larger database of potential solutions, to generate an instruction whose execution will cancel the error and minimize surprisal (Friston 2003; Hohwy et al. 2008). The process iterates until error signals are cancelled by suitable action.

This is a very basic outline of the predictive coding idea dodges a crucial question: the extent to which Bayesian formalisations actually describe neurocomputational processes rather than serving as a predictive calculus for neuroscience (Jones & Love 2011; Hohwy 2013; Clark 2012; Park & Friston 2013; Moutoussis et al. 2014). It also blurs an important distinction which is not salient to formalisations such as Bayesian theory: namely the fact that not all higher level control systems can and do smoothly cancel prediction errors generated at

³ There is an interesting debate to be had here. On the views of e.g., Damasio and Bechara affective feelings are not metacognitive but experiences produced by lower level or first order processes *associated* with metacognitive processes (such as planning and decision making). Proust refers to feelings generated by metacognitive processes. On the view proposed here the AIC metarepresents the *significance* of first order bodily information (e.g., visceral or tonic muscular state) in the context of self-relevant metacognition. It allows the subject to experience not just body state but the relevance of that body state.

lower levels. For example vision and motor control are good examples of predictive coding systems (Hohwy 2013). Often however experiences best explained as carrying information about prediction error are not cancelled by the adoption of a higher-level belief. Consider déjà vu experiences which signal mismatch between an affect of familiarity and perception of a novel scene (O'Connor & Moulin 2010). We know the scene is novel, but it still feels familiar. The point is just that the higher order belief does not always smoothly cancel prediction error. And this should be expected. Coding formats are not uniform across cognitive systems, which is why sensory and higher-level cognitive integration is such a cognitive achievement for the mind.

From our point of view what matters are the key ideas of hierarchical organization, upward referral of surprisal and top-down cancellation of error. Also crucial is the idea that the highest levels of cognitive control involve active, relatively unconstrained, exploration of solution space. This is the level at which attention can be redirected to alternative solutions and their imaginative rehearsal. Phenomena such as delusion represent a high level response to an obstinate signal of prediction error that cannot be simply cancelled from the top down. This way of thinking of the mind weds a version of predictive coding theory to insights from neuro-computational theory that treat executive systems as specialized for the resolution of problems which cannot be solved at lower levels. Thus at low levels in the hierarchy the structure of priors and errors and referral of surprisal is constrained, modularized some might say. At the so-called personal level of belief fixation predictive coding best describes the idea that those experiences which command executive resources are those which signal prediction error which cannot be resolved at lower perceptual and quasi perceptual levels. This is at least one level at which predictive coding involves active sampling of information (active inference) as well as the routine cancelling of surprisal according to a well defined prior model. The latter almost defines perception. The former, according to O'Reilly & Munakata (2000) as well as

predictive coding theorists (Spratling 2008) is definitive of executive control.

Thus most of the detection and correction of error occurs at low levels in the processing hierarchy at temporal thresholds and using coding formats that are opaque to introspection. Keeping one's balance, parsing sentences and recognizing faces are examples. We have no introspective access to the cognitive operations involved and are aware only of the outputs. This is the sense in which our mental life is tacit: automatic, hard to verbalize, and experienced as fleeting sensations that vanish quickly in the flux of experience. This is the "Unbearable Automaticity of Being" (Bargh & Chartrand 1999). However even these relatively automatic processes generate experiences of which we can become aware. The recognition of faces, for example, produces an affective response within a few hundred milliseconds. When that affective response is absent or suppressed due to malfunction a prediction is violated and the discrepancy between familiar face and lack of familiar affect is referred to higher levels of executive control to deal with the problem.

At the higher levels of cognitive control, surprisal is signalled as experience that becomes the target of executive processes. These meta-cognitive processes evolved to enable humans to reflect and deliberate to control their behaviour. The highest levels of cognitive control involve reflection, deliberation, rehearsal and evaluation of alternative courses of action and explicit reasoning. When for example a predicted affect is absent we might find ourselves in the position of a patient described by Brighetti who lost affective responses to her family and her professor. She had "identity recognition of familiar faces, associated with a lack of SCR [SCR is skin conductance response, a measure of electrodermal activity consequent on affective processing]" (Brighetti et al. 2007). In other words her predicted affective response to familiars was absent, which resulted in an experience becoming the target of higher-level control processes. Such patients sometimes produce the Capgras delusion that the familiar person has been replaced by an imposter or double. A truly florid delusion such as is sometimes seen in schizo-

phrenia might elaborate the delusional thought into an epic paranoid narrative.

The aim here is not to enter into the controversy about the explanation of the Capgras delusion but to note the role of the architecture that generates it (Young et al. 1994; Breen et al. 2001; Ellis & Lewis 2001). Higher levels of cognitive control are engaged to deal with error signals referred from lower levels in the hierarchy. Perhaps the most important level in the hierarchy for personal and social life is the level at which subjectively adequate narratives are generated to make experience intelligible and by which we communicate our experiences to others. This is the level at which delusional thoughts originate. By subjectively adequate here I merely mean “fits the experience of the subject”. At even higher levels of cognitive control we can revise and reject those subjectively adequate autobiographical narratives, replacing them with empirical theories that draw on publicly available norms of reasoning and semantic knowledge to produce objectively adequate responses to subjective experience (Gerrans 2014). Delusions are best conceptualized as higher-level responses to prediction error which, however, cannot cancel those errors. In fact as Clark (2013) points out such delusory models in effect “predict” further experiences of that type, which means that the delusion will be strengthened.

A very important point to note for the subsequent explanation of depersonalization and the Cotard delusion is that it is not the absence of affect *per se* which produces the error signal and engages higher-level cognition. Lack of affective response alone does not require a high level response unless that lack of affect is unpredicted. That is why we are not bothered by lack of response to strangers (we don’t predict it at any level in the control hierarchy) but if a new mother has no affective response to her baby the experience can be part of a syndrome of post-natal depression.

The example of post-natal depression allows us to make another important point about the relationship between predicted affect and psychosis. Mothers most vulnerable to post-natal depression are those who had powerful

positive expectations of motherhood and the bond with the infant. When that bond does not materialize for some reason they are confronted with a distressing lack of predicted affective response. Sometimes this will produce a kind of Capgras delusion regarding the baby. The mother might say that the baby has been replaced or is an alien (Brockington & Kumar 1982). Interestingly, and tellingly, if the mother is also extremely anxious the condition can be even more serious. Anxious attention to the experience tends to magnify the problem.

This role for anxiety is nicely elucidated by the predictive coding framework. Formal considerations aside, the concept of predictive coding places a huge emphasis on the signaling of error. This means that incoming information must be compared to a prediction and the difference computed and referred to a control system. At higher levels those error signals take the form of experiences. These experiences are often imprecise and opaque since they are produced by lower level systems that encode information in different formats to those used by explicit metarepresentational capacities. They also compete for metarepresentational resources among the constant flux of experiences that engage attention. Thus they create a problem of working out for any experience how much is signal and how much is noise.

It is very important for high-level cognition to be targeted as precisely as possible for only as long as required. Thus any vagueness in experience needs to be resolved. Attention is the process which solves this problem. Hohwy (2012, p. 1; my emphasis) makes the point for perceptual inference but it applies in general:

conscious perception can be seen as the upshot of prediction error minimization and *attention* as the optimization of precision expectations during such perceptual inference.

Clark (2013, p. 190) makes a similar point:

Attention, if this is correct, is simply one means by which certain error-unit responses are given increased weight, hence

becoming more apt to drive learning and plasticity, and to engage compensatory action.

The point is that attention is directed to error signals in order to make them more precise by increasing the signal to noise ratio. Attention amplifies the signal and maintains it while higher-level systems try and interpret the experience and manage appropriate responses. If the response works the error signal is cancelled and attention can be directed elsewhere.

Within this framework we can make an observation about anxiety that can be overlooked by approaches that concentrate on the arousal, hypervigilance or the associated beliefs concerning threat or danger. These approaches de-emphasise a crucial element. That is uncertainty. Anxiety is an adaptive mechanism that primes the organism cognitively and physiologically to resolve uncertainty. Thus, if a prediction cannot be verified, or an error signal disambiguated, anxiety in this sense will result. Of course what we call pathological anxiety is the dysfunctional activation and maintenance of these mechanisms. The point is that someone who is anxious in this way will continue to misallocate attentional, cognitive and physiological resources to experiences. Another point about anxiety is that, in pathological cases, action does not cancel the signal or the dysfunctional allocation of resources to it. This may be why the role of anxiety in depersonalisation is not straightforward. Some recent studies have not found a strong correlation between anxiety and depersonalisation (e.g., [Medford 2012](#)). However the scales used to measure anxiety give a score that sums scores for self-report of feelings, behaviour and cognition. The suggestion here is that what really matters is the allocation of attention to signals which cannot be resolved, perhaps because they are intrinsically noisy, ambiguous or have insufficient information. It is also important that the patient cannot resolve the uncertainty by revising the predictive model that generates it since that is usually maintained low in the predictive hierarchy by mechanisms that are not accessible. The person with Capgras delusion, for example, automatically

predicts affective response to familiar faces and when it goes missing there is nothing she can do to revise that prediction. Instead she is confronted with an anomalous experience, which automatically captures attention. Similarly with depression. Loss of affective response is not something that can be restored from the top down.

In some cases of post-natal depression all these factors seem to be operative. The mother expected to bond with the infant but in fact perhaps birth was traumatic, the baby did not attach straightaway, and the mother needed more support and reassurance than she received. She was left distressed and unable to cope which made bonding and attachment even more difficult. This would be bad enough but if the mother had a strong prior expectation that motherhood would be straightforwardly rewarding a prediction is violated. If the mother is also anxious she will attend intensively to the resultant experience of absent affect, but she will encounter only further feelings of emptiness and panic. The presence of the baby and the expectations of family and friend only compound the sense that she is not feeling what she should be feeling. What happens next depends on context and support but it is not really surprising, especially given the relationship between massive hormonal fluctuation and emotional regulation, that in some cases new mothers develop psychotic symptoms ([Spinelli 2009](#)).

7 Depersonalisation

Depersonalisation Disorder (DPD) is characterized by “alteration in the perception or experience of the self so that one feels detached from and as if one is an outside observer of one’s own mental processes” ([American Psychiatric Association 2000](#)). Critchley points out that DPD is often accompanied by alexithymia, a condition in which conscious awareness of emotional states is compromised or absent. This is consistent with findings summarized by Medford that “de-affectualisation”, a reduction or absence of affective response, presents as a core feature of clinical cases. Depersonalisation is a separate disorder to derealisation (the feeling that the world is inanimate or unreal) but derealisation

is often an important aspect of depersonalisation. Indeed, as Medford describes their relationship, depersonalisation can sometime be a response to derealisation (Sierra et al. 2002; Hunter et al. 2004).

Seth et al. (2011, p. 9; my emphasis) summarize a range of findings about DPD as follows: “In short, DPD can be summarized as a psychiatric condition marked by the selective diminution of the *subjective reality* of the self and world”. They explain this diminution as the result of the loss of “sense of presence”, the feeling of being engaged in experience. This is what they mean by subjective reality: the condition is not like an hallucination or delusion in which objective reality is misrepresented by faulty perception or belief fixation. In fact the patient correct represents “objective reality” but loses the sense of herself as the subject of experience.

In the attempt to explain the loss of the sense of presence cognitive neuroscience has developed a theoretical picture that considerably augments older theories. On those older theories DPD represented a suppression or inhibition of emotion as a response to trauma or distress. On this view DPD activates mechanisms which might in other circumstances be adaptive. For example, if the subject of violent attack deactivated those mechanisms which produce the experience of distress that would qualify as an adaptive response to trauma. Of course such a response is only adaptive in the short term. Inability to feel distress might also reduce avoidance behavior with disastrous consequences.

It seems that the deactivation is accomplished by inhibitory activity in the Ventrolateral Prefrontal Cortex (VLPFC). The VLPFC is a structure which plays a crucial role in the regulation of affective feeling, especially as part of a process of reappraisal (Füstös et al. 2013). The adaptive aspect here is that it allows the subject to redirect attention and divert cognitive resources to alternative interpretations of self-relevance and response behaviour by inhibiting an experience that would otherwise monopolise cognition. This role has been tested in tasks which involve the top down regulation of negative affect but, as Medford says, “In DPD such suppression is apparently involuntary (and

largely resistant to volitional control), but it is reasonable to suppose that this will nevertheless engage similar inhibitory networks” (2012, p. 142). Thus the patient with DPD experiences the result of *involuntary* deactivation of systems that produce the bodily experience of emotion.

These ideas are consistent with the evidence from cognitive neuroscience about other primary neural correlates of DPD. *Hyperactivity* in VLPFC leads to *hypoactivity* in the Anterior Insular Cortex (AIC). That reduced activity in the AIC produces the loss of a sense of presence. This hypothesis results from findings that it has a primary role in higher order representation of interoceptive (visceral, autonomic, bodily) states. It generates the bodily feelings that signal how we are faring in the world moment to moment consequent on affective processing. Activity in the AIC produces what Damasio called the “core self” and what Critchley calls “the sense of presence”. As Critchley says,

evidence from a variety of sources converges to suggest a representation of autonomic and visceral responses within anterior insula cortex, where, particularly on the right side, this information is accessible to conscious awareness, influencing emotional feelings (2005, p. 162).

When Damasio made his contributions to the neurophilosophy of emotions and self-representation the computational theory in the field was less developed so that we can now make some additional observations about the role of the AIC.

To do so we first reiterate the distinction between being able to sense body state, which is the phenomenon baptized by Damasio *interoception*, (to distinguish it from *exteroception* [perception of the external world]), and sensing states of a self. The distinction is a subtle one of course but we can approach it intuitively by noting that there is a crucial difference between being able to sense heart rate, blood pressure or temperature as part of an illness and as part of an emotional episode. We observed earlier that the second kind of awareness is the one we describe as self-awareness in virtue of the fact that

it reflects affective processing rather than pure bodily regulation. There is a difference in feeling state caused by raises in blood pressure generated by walking up stairs and by heated argument. This is so even though heart rate is heart rate, however caused. But the point of affective processing, as we saw, is to assess the self-relevance of unpredicted changes in things like heart rate and to indicate to the subject how and why they might matter in the cognitive context.

The experiential differences between heart rate *per se* and heart rate consequent on affective processing can be explained in terms of the principle of hierarchical computational organization, reflected in cortical organization (Craig 2009, 2010; Dunn et al. 2010). The insular cortex is hierarchically organized to map body state at different levels of abstraction and integration. Posterior sections map body state directly and integrate those representations to coordinate *reflexive* regulatory functions. Thus the Posterior Insular Cortex (PIC), for example, represents things like blood pressure and departures from homeostasis and integrates that information to enable reflexive regulatory processing. More anterior regions re-represent and integrate this information in formats available for higher levels of cognitive control. If we sense raised blood pressure the PIC is primary in the representation of that information. When, however, we are deciding how to respond, we need to integrate that information with current and long term goals, representations of contextual information, memory, planning and inference. We may have to inhibit or reprogram automatic behavioral tendencies (not punch the boss) and perhaps reappraise the situation. Thus we need a way, not just to feel raised blood pressure, but to *feel its significance* in order to program a suitable response. This is the role of the AIC.

This explains a recent finding which seems paradoxical on the “somatic” James-Lange view of emotions revived by Damasio. On that view emotions are representations of body state *simpliciter*. The feeling of fear is the feeling of being primed to take avoidance action, for example. Michal and collaborators compared the “interoceptive accuracy”, that is ability of patients to judge body state (using heart rate as a

proxy), of patients with DPD to normal patients. Strikingly they found that “[there] was no correlation of the severity of ‘anomalous body experiences’ and depersonalization with measures of interoceptive accuracy.” They explained this finding as follows: “[The] findings highlight a striking discrepancy of normal interoception with overwhelming experiences of disembodiment in DPD. *This may reflect difficulties of DPD patients to integrate their visceral and bodily perceptions into a sense of their selves*” (Michal & Reuchlein 2014, p. 1; my emphasis).

The AIC can only integrate currently available bodily feeling. As Craig says, it “represents the sentient self at one moment of time [and] provides the basis for the continuity of subjective emotional awareness in a finite present” (2009, p. 67). However we can extend the temporal range of information represented by those feelings by integrating them with representation of past and future episodes of experience and/or semantic knowledge. Simulations involved in planning and episodic memory are associated with activation of the AIC to provide sense of extended self. In other words it is the integration of the metarepresentations of body state produced by the AIC with representations of episodes of a temporally extended autobiography that produces the feeling that we are a self with a past and future, rather than a series of disconnected selves, moment to moment.

Nothing in what I have said refutes skepticism about the self, or episodic theories of first person experience (Strawson 2004). It is in fact consistent with the idea that experience is a series of episodes. Whether we feel those episodes are ours depends on how they are integrated. There is no suggestion that everyone integrates them the same way or that integration evokes an equally strong sense of presence in each person. All I have suggested is that there are mechanisms which can create self-awareness moment to moment and mechanisms which integrate those moments of self-awareness with higher level forms of cognitive control that represent past and future actions and outcomes in order for the organism to assess the self-relevance of actual and potential actions. The ex-

planation of awareness of self-relevance in different contexts is a sufficient explanation of the phenomenon of self-awareness that was our initial quarry.

Craig adds a subtle but important qualification to this account. He (and others) remind us that if the predictive coding account of the mind is correct then we are never directly aware of objects, including the body (Craig 2009, 2010; Seth et al. 2011; Garfinkel & Critchley 2013). Rather representations of objects are computed on the basis of discrepancy between their predicted informational effects on us and actual incoming information. It is fluctuations and discrepancies measured against expectations computed at different levels in the control hierarchy that determine the information that becomes consciously available. “An *expected* event does not need to be explicitly represented or communicated to higher cortical areas which have processed all of its relevant features prior to its occurrence” (Bubic et al. 2010, p. 10; Clark 2013, p. 199; my emphasis.)

The same should be true of neural activation in the AIC, and hence of moments of self-awareness. We are aware of what is relevant to us via unpredicted changes in bodily feeling consequent on affective processing.

This latter feature is the key to understanding the link between “de-affectualisation”, as Medford called it, and depersonalization (Medford 2012). It is not the fact that affect is suppressed that matters, but that affect which was predicted to occur does not in virtue of the *involuntary* inhibition of the AIC by the VLPFC. When people engage in voluntary or effortful inhibition of affect they do not feel depersonalized. We noted earlier the role of expectation in post-natal depression, but there the expectation is of affective response to a specific object, a baby. In depersonalization it seems that almost all expected affective feelings are absent because of hyperactivity in the VLPFC.

The predictive coding framework also allows us to finesse explanations of the role of anxiety in the experience of derealisation. We noted that Cotard described anxiety as part of

the aetiology of the depersonalization experience in Cotard delusion. Medford, in an early discussion of DPD, also postulated a role for anxiety in order to explain an apparent paradox of DPD: the distress experienced by the patient at the absence of affective response. It is not merely that the patient has no emotions, but, as a patient of Medford’s said, “I don’t have any emotions—it makes me so unhappy.” Medford (2012) pointed out that this is only slightly paradoxical: the distress is at the lack of *internal* affect, the inability to feel rather than at the derealisation of the external world. Medford related this specifically to the anxiety component of the syndrome. The patient expects that the world will induce positive affect but when it does not an expectation is violated and the patient anxiously attends to that absence of affect. On this view highly anxious patients are hyperattentive to their experience and encounter, not the normal bodily experience, which represents how they are faring in the world, but a strange absence of such experience, in combination with intact exteroception which tells them that the world is unchanged (Paulus & Stein 2010; Garfinkel & Critchley 2013; Seth 2013; Terasawa et al. 2013). Their problem is that they no longer feel the relevance of changes in their own bodies and the world to themselves and this inability to feel the world increases their anxiety. Medford quotes an earlier theorist (Ackner 1954) who noted “increased responsiveness for anxiety of internal origin, whereas that of external origin [is] reduced” (Medford 2012, p. 141).

This perhaps explains the differences in casual aetiology between depersonalisation arising in the Cotard syndrome and in DPD. In the Cotard syndrome something is amiss with the mechanisms that appraise perceptual and interoceptive information for self-relevance. The AIC is not getting any information from affective systems to integrate and relay to higher order cognition. Thus felt significance disappears. When the depressive patient then focuses on her experience she feels alienated from the world and depersonalised. In the case of DPD it appears that the AIC is

hypoactive for another reason: its activity is inhibited by the VLPFC.

In both cases the patient attends to her experience and tries to interpret it in order to respond. This is consistent with the role postulated by predictive coding theories for attention: the attempt to interpret and sharpen the informational content of a signal by improving the signal to noise ratio. Unfortunately an increase in attention does not provide any increase in precision, it only makes the absence of predicted response more salient. Since those predictions are, in effect, representations of expected self-relevance that normally provide the experience of self-awareness, the patient concludes that the self does not exist. After all, the information necessary to generate self-awareness is still in place. The body, the world and first order representations of their interaction are all represented in experience. What is lost is a sense of the significance of those interactions for the body that mediates them.

The explanation has become complicated so at this point it is useful to situate it in terms of the conceptual architecture (points (i)-(iii) below) outlined in the introduction. On this view DPD arises in the following way as a personal level response to the absence of predicted affective experience.

- i. Appraisal systems normally represent the significance of information for the organism. The primary way of experiencing the result of those appraisals is via activation in the AIC. This is because the AIC is specialised for informing the subject, via bodily experience, of the affective significance of its encounters with the world. These experiences are not the same as experience of body state *per se* but the emotional significance of that body state.
- ii. Those experiences can be rehearsed offline in planning and deliberation to extend the temporal horizon of affective experience. We feel like temporally integrated selves because memory and prospection have affective significance.
- iii. Predictive coding architecture has the effect of making discrepancy between anticipated and actual affective feeling highly salient.
- iv. In DPD activity in the AIC is inhibited most likely as a result of the involuntary activity of the VLPFC.
- v. Consequently the patient has normal perceptual and sensory responses to the world but those responses are not integrated into a bodily representation which informs her of their significance. The world feels derealised or as Medford puts it de-affectualised
- vi. However, given the way predictive coding works, the patient actually has a model of the world that predicts activity in the AIC as a result of her perceptual encounters. Thus absence of AIC-produced experience is a prediction error that drives metacognitive responses.
- vii. Those responses include increased attention (driven by sub personal mechanisms of resource allocation) to the experience itself as the patient tries to extract further information from it. However, being produced by subpersonal mechanisms the experience is both intractable and inscrutable.
- viii. Highly anxious people cannot divert attention from the experience, since anxiety is driven by the need to resolve uncertainty. But the experience is inexplicable and irresolvable.
- ix. The patient's personal level interpretation of the experience is of depersonalisation "it feels like it is not happening to me". The interpretation is not a direct report of the experience, which I have argued is more like a total deaffectualisation. It amplifies it.
- x. However the form that amplification takes, depersonalisation, is explained by the role such experiences have in creating the normal sense of being a self. We feel we are selves precisely because the significance of the world for our organismic goals is normally computed by appraisal systems and represented in characteristic forms of bodily experience.

8 Anatomy of an avatar

Thomas Metzinger has argued that the persisting self is neither an illusion (in the sense of a perceptual experience whose content is incorrect) nor a genuine entity in the sense of an object existing outside the mind like a body or a neural state. Instead the self is a creature of experience itself, a phenomenal representation constructed by the brain to control the body. This representation is in effect an avatar that unifies experiences of ownership (the sense of the integrity of bodily boundaries), perspective on experience (which I have not talked about in this essay), and selfhood (“a single coherent and temporally stable phenomenal subject”). An especially attractive aspect of Metzinger’s view is that he treats the nature of the avatar as an empirical matter so that our understanding of its properties can be refined in the face of further discoveries.

Metzinger’s view nicely captures what is right and wrong in the illusory view of the self. The illusory view is correct that the self is not an object to be experienced in the same way as we experience perceptual or somatic objects. The self is a way of experiencing the interaction of the body and the world. It is a creature of experience, constructed by the brain to navigate the organism through the world. The self exists as a virtual phenomenal entity in virtue of the integrative processes that create and sustain it.

The Fat Controller view of the self also has some of the picture correct. Self-awareness is needed for higher order cognitive control to integrate and organise experience moment to moment and to assimilate those experiences to an ongoing autobiography for longer-term cognitive control. However there is no single cognitive process with an identifiable neural substrate that represents an organiser/narrator. Also, and this is where Metzinger is correct, there is a genuine experience of being a person in control, but this experience is the experience of integration itself, which suggests that it is a process which can disintegrate and degrade in different ways and to different degrees. It also suggests, although I have not discussed it here, that experience of the self is a prefrontal achievement

since prefrontal structures are specialised for “large world” integrative processing (the orchestration of synchronised activity across widely distributed brain areas).

The Embodied Self view is of course very close to the one I have discussed here. I have argued that a particular type of bodily feeling is what goes awry in depersonalisation and hence that those feelings produce the experience of the self. While this is correct, we need to recall that Damasio distinguished between the “core self”, which is very close to the phenomenon I have described, and the autobiographical self. Sometimes he treats the autobiographical self as a more abstract or narrative construct. I have tried to show that the integration of the core self with the autobiographical self comes, as it were, for free, given the automatic links between affective processing and the processes which construct the autobiographical self. It is impossible to rehearse episodes of one’s autobiography without a sense of presence—unless, of course, one has DPD or the Cotard delusion. But those cases demonstrate the component structure of the avatar.

Finally, the narrative view captures the crucial role of temporal integration in the experience of the self. But the self is not *just* a fictional protagonist in the brain’s stories (though it is that). The specialised simulation mechanisms that generate the actual and potential autobiographies automatically integrate each episode with affective feeling. That feeling allows us to experience in the process of recollection, imagination or narration the significance of each episode to our unique organismic trajectory. That, and the ability to incorporate and act on those feelings, is all the selfhood anyone needs.

References

- Ackner, B. (1954). Depersonalization: I. Aetiology and phenomenology. *British Journal of Psychiatry*, 100 (421), 838-853. [10.1192/bjp.100.421.838](https://doi.org/10.1192/bjp.100.421.838)
- Adolphs, R. (2010). What does the amygdala contribute to social cognition? *Annals of the New York Academy of Sciences*, 1191 (1), 42-61. [10.1111/j.1749-6632.2010.05445.x](https://doi.org/10.1111/j.1749-6632.2010.05445.x)
- Adolphs, R., Baron-Cohen, S. & Tranel, D. (2002). Impaired recognition of social emotions following amygdala damage. *Journal of Cognitive Neuroscience*, 14 (8), 1264-1274. [10.1162/089892902760807258](https://doi.org/10.1162/089892902760807258)
- Allport, G. W. (1961). *Pattern and growth in personality*. New York, NY: Holt, Rinehart & Winston.
- American Psychiatric Association, (2000). *Diagnostic and statistical manual of mental disorders: DSM-IV-TR*. Washington DC: American Psychiatric Association.
- Bargh, J. A. & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist*, 54 (7), 462-479. [10.1037/0003-066X.54.7.462](https://doi.org/10.1037/0003-066X.54.7.462)
- Bechara, A. & Damasio, A. R. (2005). The somatic marker hypothesis: A neural theory of economic decision. *Games and Economic Behavior*, 52 (2), 336-372. [10.1016/j.geb.2004.06.010](https://doi.org/10.1016/j.geb.2004.06.010)
- Blanke, O. & Metzinger, T. (2009). Full-body illusions and minimal phenomenal selfhood. *Trends in Cognitive Sciences*, 13 (1), 7-13. [10.1016/j.tics.2008.10.003](https://doi.org/10.1016/j.tics.2008.10.003)
- Breen, N., Coltheart, M. & Caine, D. (2001). A two-way window on face recognition. *Trends in Cognitive Sciences*, 5 (6), 234-235. [10.1016/S1364-6613\(00\)01659-4](https://doi.org/10.1016/S1364-6613(00)01659-4)
- Brighetti, G., Bonifacci, P., Borlimi, R. & Ottaviani, C. (2007). "Far from the heart far from the eye": Evidence from the Capgras delusion. *Cognitive Neuropsychiatry*, 189 (197), 12-3. [10.1080/13546800600892183](https://doi.org/10.1080/13546800600892183)
- Brockington, I. F. & Kumar, R. (1982). *Motherhood and mental illness*. London, UK: Academic Press.
- Broyd, S. J., Demanuele, C. & , (2009). Default-mode brain dysfunction in mental disorders: A systematic review. *Neuroscience and Biobehavioral Reviews*, 33 (3), 279-296. [10.1016/j.neubiorev.2008.09.002](https://doi.org/10.1016/j.neubiorev.2008.09.002)
- Bubic, A., Von Cramon, D. Y. & Schubotz, R. I. (2010). Prediction, cognition and the brain. *Frontiers in Human Neuroscience*, 4 (25), 1-15.
- Buckner, R. L., Andrews-Hanna, J. R. & , (2008). The brain's default network: Anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences*, 1124, 1-38. [10.1196/annals.1440.011](https://doi.org/10.1196/annals.1440.011)
- Clark, A. (2012). Dreaming the whole cat: Generative models, predictive processing, and the enactivist conception of perceptual experience. *Mind*, 121 (483), 753-771. [10.1093/mind/fzs106](https://doi.org/10.1093/mind/fzs106)
- (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36 (3), 181-204. [10.1017/S0140525X12000477](https://doi.org/10.1017/S0140525X12000477)
- Cotard, J. (1880). Du délire hypocondriaque dans une forme grave de la mélancolie anxieuse. *Annales Médico-Psychologiques*, 38, 168-170.
- (1882). Du délire des négations. *Archives de Neurologie*, 4, 282-295.
- (1884). Perte de la vision mentale dans le mélancolie anxieuse. *Archives de Neurologie*, 7, 289-295.
- (1891). *Études sur les maladies cérébrales et mentales*. Paris, FR: Baillière.
- Craig, A. D. (2009). How do you feel now? The anterior insula and human awareness. *Nature Reviews Neuroscience*, 10 (1), 59-70. [10.1038/nrn2555](https://doi.org/10.1038/nrn2555)
- (2010). The sentient self. *Brain Structure and Function*, 214 (5), 563-577. [10.1007/s00429-010-0248-y](https://doi.org/10.1007/s00429-010-0248-y)
- Critchley, H. D. (2005). Neural mechanisms of autonomic, affective, and cognitive integration. *Journal of Comparative Neurology*, 493 (1), 154-166. [10.1002/cne.20749](https://doi.org/10.1002/cne.20749)
- Currie, G. & Jureidini, J. (2004). Narrative and coherence. *Mind and Language*, 19 (4), 409-427. [10.1111/j.0268-1064.2004.00266.x](https://doi.org/10.1111/j.0268-1064.2004.00266.x)
- Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York, UK: Putnam.
- Debruyne, H., Portzky, M., van den Eynde, F. & Aude-naert, K. (2009). Cotard's syndrome: A review. *Current Psychiatry Reports*, 11 (3), 197-202. [10.1007/s11920-009-0031-z](https://doi.org/10.1007/s11920-009-0031-z)
- Dunn, B. D., Galton, H. C., Morgan, R., Evans, D., Oliver, C., Meyer, M., Cusack, R., Lawrence, A. D. & Dalgleish, T. (2010). Listening to your heart: How interoception shapes emotion experience and intuitive decision making. *Psychological Science*, 21 (12), 1835-1844. [10.1177/0956797610389191](https://doi.org/10.1177/0956797610389191)
- Ellis, H. D. & Lewis, M. B. (2001). Capgras delusion: A window on face recognition. *Trends in Cognitive Sciences*, 5 (4), 149-156. [10.1016/S1364-6613\(00\)01620-X](https://doi.org/10.1016/S1364-6613(00)01620-X)
- Enoch, M. D. & Trethowan, W. H. (1991). *Uncommon psychiatric syndromes*. Oxford, UK: Butterworth-Heinemann.
- Fair, D. A., Cohen, A. L. & , (2008). The maturing architecture of the brain's default network. *Proceedings of the National Academy of Sciences of the United States of America*, 105 (10), 4028-4032. [10.1073/pnas.0800376105](https://doi.org/10.1073/pnas.0800376105)

- Friston, K. (2003). Learning and inference in the brain. *Neural Networks*, 16 (9), 1325-1352. [10.1016/j.neunet.2003.06.005](https://doi.org/10.1016/j.neunet.2003.06.005)
- Füstös, J., Gramann, K., Herbert, B. M. & Pollatos, O. (2013). On the embodiment of emotion regulation: Interoceptive awareness facilitates reappraisal. *Social Cognitive and Affective Neuroscience*, 8 (8), 911-917. [10.1093/scan/nss089](https://doi.org/10.1093/scan/nss089)
- Garfinkel, S. N. & Critchley, H. D. (2013). Interoception, emotion and brain: New insights link internal physiology to social behavior. *Social Cognitive and Affective Neuroscience*, 8 (3), 231-234. [10.1093/scan/nss140](https://doi.org/10.1093/scan/nss140)
- Gerrans, P. (2000). A one-stage explanation of the Cotard delusion. *Philosophy, Psychiatry, and Psychology*, 9 (1), 47-53. [10.1353/ppp.2003.0007](https://doi.org/10.1353/ppp.2003.0007)
- (2001). Delusions as performance failures. *Cognitive Neuropsychiatry*, 6 (3), 161-173.
- (2014). *The measure of madness: Philosophy of mind, cognitive neuroscience, and delusional thought*. Cambridge, MA: MIT Press.
- Gerrans, P. & Kennett, J. (2010). Neurosentimentalism and moral agency. *Mind*, 119 (475), 585-614. [10.1093/mind/fzq037](https://doi.org/10.1093/mind/fzq037)
- Gilboa, A. (2004). Autobiographical and episodic memory one and the same? Evidence from prefrontal activation in neuroimaging studies. *Neuropsychologia*, 42 (10), 1336-1349. [10.1016/j.neuropsychologia.2004.02.014](https://doi.org/10.1016/j.neuropsychologia.2004.02.014)
- Goldie, P. (2011). Life, fiction, and narrative. In N. Carroll & J. Gibson (Eds.) *Narrative, emotion, and insight* (pp. 8-22). University Park, PA: Pennsylvania State University Press.
- Gusnard, D. A., Akbudak, E., Shulman, G. L. & Raichle, M. E. (2001). Medial prefrontal cortex and self-referential mental activity: Relation to a default mode of brain function. *Proceedings of the National Academy of Sciences of the United States of America*, 98 (7), 4259-4264. [10.1073/pnas.071043098](https://doi.org/10.1073/pnas.071043098)
- Halligan, P. W. & Marshall, J. C. (1996). *Method in madness: Case studies in cognitive neuropsychiatry*. Hove, UK: Psychology Press.
- Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology*, 3 (96), 1-14. [10.3389/fpsyg.2012.00096](https://doi.org/10.3389/fpsyg.2012.00096)
- (2013). *The predictive mind*. Oxford, UK: Oxford University Press.
- Hohwy, J., Roepstorff, A. & Friston, K. (2008). Predictive coding explains binocular rivalry: A review. *Cognition*, 108 (3), 687-701. [10.1016/j.cognition.2008.05.010](https://doi.org/10.1016/j.cognition.2008.05.010)
- Hunter, E. C., Sierra, M. & David, A. S. (2004). The epidemiology of depersonalisation and derealisation. *Social Psychiatry and Psychiatric Epidemiology*, 39 (1), 9-18. [10.1007/s00127-004-0701-4](https://doi.org/10.1007/s00127-004-0701-4)
- Jones, M. & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, 34 (4), 169-188. [10.1017/S0140525X10003134](https://doi.org/10.1017/S0140525X10003134)
- Jureidini, J. (2012). Explanations and unexplanations: Restoring meaning to psychiatry. *Australia and New Zealand Journal of Psychiatry*, 46 (3), 188-191. [10.1177/0004867412437347](https://doi.org/10.1177/0004867412437347)
- Kenny, A. (1963). *Action, emotion & will*. London, UK: Routledge & Kegan Paul.
- Koenigs, M. & Grafman, J. (2009). The functional neuroanatomy of depression: Distinct roles for ventromedial and dorsolateral prefrontal cortex. *Behavioral Brain Research*, 201 (2), 239-243. [10.1016/j.bbr.2009.03.004](https://doi.org/10.1016/j.bbr.2009.03.004)
- Medford, N. (2012). Emotion and the unreal self: Depersonalization disorder and de-affectualization. *Emotion Review*, 4 (2), 139-144. [10.1177/1754073911430135](https://doi.org/10.1177/1754073911430135)
- Metzinger, T. (2003). *Being no one: The self-odel theory of subjectivity*. Cambridge, MA: MIT Press.
- (2011). The no-self alternative. In S. Gallagher (Ed.) *The Oxford Handbook of the Self* (pp. 279-296). Oxford, UK: Oxford University Press.
- Michal, M. & Reuchlein, B. (2014). Striking discrepancy of anomalous body experiences with normal interoceptive accuracy in depersonalization-derealization disorder. *PLoS One*, 9 (2), e89823-e89823. [10.1371/journal.pone.0089823](https://doi.org/10.1371/journal.pone.0089823)
- Moutoussis, M., Fearon, P., El-Deredy, W., Dolan, R. J. & Friston, K. J. (2014). Bayesian inferences about the self (and others): A review. *Consciousness and cognition*, 25 (100), 67-76. [10.1016/j.concog.2014.01.009](https://doi.org/10.1016/j.concog.2014.01.009)
- Murphy, D. (2006). *Psychiatry in the scientific image*. Cambridge, UK: MIT Press.
- N'Diaye, K., Sander, D. & Vuilleumier, P. (2009). Self-relevance processing in the human amygdala: Gaze direction, facial expression, and emotion intensity. *Emotion*, 9 (6), 798. [10.1037/a0017845](https://doi.org/10.1037/a0017845)
- Ochsner, K. N., Bunge, S. A. & , (2002). Rethinking feelings: An fMRI study of the cognitive regulation of emotion. *Journal of Cognitive Neuroscience*, 14 (8), 1215-1229. [10.1162/089892902760807212](https://doi.org/10.1162/089892902760807212)
- O'Connor, A. R. & Moulin, C. J. (2010). Recognition without identification, erroneous familiarity, and déjà

- vu. *Current Psychiatry Reports*, 12 (3), 165-173. [10.1007/s11920-010-0119-5](#)
- O'Reilly, R. C. & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. Cambridge, MA: MIT Press.
- Park, H. J. & Friston, K. (2013). Structural and functional brain networks: From connections to cognition. *Science*, 342 (6158), 1238411-1238411. [10.1126/science.1238411](#)
- Paulus, M. P. & Stein, M. B. (2010). Interoception in anxiety and depression. *Brain Structure and Function*, 214 (5-6), 451-463. [10.1007/s00429-010-0258-9](#)
- Prinz, J. J. (2004). *Gut reactions: A perceptual theory of emotion*. Oxford, UK: Oxford University Press.
- Proust, J. (2013). *The philosophy of metacognition: Mental agency and self-awareness*. Oxford, UK: Oxford University Press.
- Sander, D., Grafman, J. & Zalla, T. (2003). The human amygdala: An evolved system for relevance detection. *Reviews in the Neurosciences*, 14 (4), 303-316.
- Sander, D., Grandjean, D. & Scherer, K. R. (2005). A systems approach to appraisal mechanisms in emotion. *Neural Networks*, 18 (4), 317-352. [10.1016/j.neunet.2005.03.001](#)
- Scherer, K. R. (2004). Feelings integrate the central representation of appraisal-driven response organization in emotion. In A. S. R. Manstead, N. Frijda & A. Fischer (Eds.) *Feelings and emotions: The Amsterdam Symposium* (pp. 136-157). Cambridge, UK: Cambridge University Press.
- Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, 17 (11), 565-573. [10.1016/j.tics.2013.09.007](#)
- Seth, A. K., Suzuki, K. & Critchley, H. D. (2011). An interoceptive predictive coding model of conscious presence. *Frontiers in Psychology*, 2 (395), 1-16. [10.3389/fpsyg.2011.00395](#)
- Sierra, M., Lopera, F., Lambert, M. V., Phillips, M. L. & David, A. S. (2002). Separating depersonalisation and derealisation: The relevance of the "lesion method". *Journal of Neurology, Neurosurgery & Psychiatry*, 72 (4), 530-532. [10.1136/jnnp.72.4.530](#)
- Solomon, R. C. (1976). *The passions: Emotions and the meaning of life*. Indianapolis, ID: Hackett.
- (1993). The philosophy of emotions. In J. M. Haviland & M. Lewis (Eds.) *Handbook of emotions* (pp. 3-15). New York, NY: Guildford Press.
- Spinelli, M. (2009). Postpartum psychosis: Detection of risk and management. *American Journal of Psychiatry*, 166 (4), 405-408. [10.1176/appi.ajp.2008.08121899](#)
- Spratling, M. W. (2008). Reconciling predictive coding and biased competition models of cortical function. *Frontiers in Computational Neuroscience*, 2 (4), 1-8. [10.3389/neuro.10.004.2008](#)
- Strawson, G. (2004). Against narrativity. *Ratio*, 17 (4), 428-452. [10.1111/j.1467-9329.2004.00264.x](#)
- Terasawa, Y., Shibata, M., Moriguchi, Y. & Umeda, S. (2013). Anterior insular cortex mediates bodily sensibility and social anxiety. *Social Cognitive and Affective Neuroscience*, 8 (3), 259-266. [10.1093/scan/nss108](#)
- Tomkins, S. S. (1962). *Affect, imagery, consciousness vol. 1: The positive affects*. New York, NY: Springer.
- (1991). *Affect, imagery, consciousness vol. 3: The negative affects: Anger and fear*. New York, NY: Springer.
- Young, A. W., Leafhead, K. M. & Szulecka, T. K. (1994). The Capgras and Cotard delusions. *Psychopathology*, 27 (3-5), 226-231. [10.1159/000284874](#)

Memory for Prediction Error Minimization: From Depersonalization to the Delusion of Non-Existence

A Commentary on Philip Gerrans

Ying-Tung Lin

Depersonalization is an essential step in the development of the Cotard delusion. Based on [Philip Gerrans'](#) account ([this collection](#)), which is an integration of the appraisal theory, the simulation theory, and the predictive coding framework, this commentary aims to argue that the role of memory systems is to update the knowledge of prior probability required for successful predictions. This view of memory systems under the predictive coding framework provides an explanation of how experience is related to the construction of mental autobiographies, how anomalous experience can lead to delusions, and thus how the Cotard delusion arises from depersonalization.

Keywords

Affective processing | Cotard delusion | Depersonalization | Memory | Narrative | Predictive coding framework | Self-awareness | Simulation model

Commentator

[Ying-Tung Lin](#)

liningtung@gmail.com

國立陽明大學

National Yang-Ming University
Taipei, Taiwan

Target Author

[Philip Gerrans](#)

philip.gerrans@adelaide.edu.au

University of Adelaide
Adelaide, Australia

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

In [Le Délire de Négation](#) (1897), Jules Séglas considers depersonalization to be an essential step in the development of the Cotard delusion¹

¹ According to [Berrios & Luque \(1995\)](#), the English translation of “le délire des négations”—a term first introduced by the French neurologist, Jules Cotard (1840–1889)—only conveys a part of what it means: “Délire is not a state of delirium or organic confusion (in French, *délire aigu* and *confusion mentale*) or a delusion (in French, *idée* or *thème délirante*)—it is more like a syndrome that may in-

(CD; as cited in [Debruyne 2009](#); [Gerrans 2002](#)), and *prima facie* the two states share a number of characteristics: Patients suffering from the former feel *as if* they are dead or do not exist,

clude symptoms from the intellectual, emotional, or volitional spheres” (p. 219). The original French concept of “délire” fits better with Gerrans’ account of the Cotard delusion, in which the Cotard delusion does not merely concern beliefs of denial, but also anomalous affective processing.

whereas those who suffer from the latter sincerely believe and experience this state. However, the central characteristics of these disorders are distinct. Patients describe the experience of depersonalization as follows:

It's really weird. It's sort of like I'm here, but I'm really not here and that I kind of stepped out of myself, like a ghost... I feel really light, you know. I feel kind of empty and light, like I'm going to float away... Sometimes I really look at myself that way... It's kind of a cold eerie feeling. I'm just totally numbed by it. (Cited in [Steinberg 1995](#))

The emotional part of my brain is dead. My feelings are peculiar, I feel dead. Whereas things worried me nothing does now. My husband is there but he is part of the furniture. I don't feel I can worry. All my emotions are blunted. ([Shorvon 1946](#), p. 783)

As illustrated in these subjective descriptions, depersonalization is characterized by a loss of the sense of presence ([Critchley 2005](#)) or an increased “sense of detachment”—the “[e]xperience of unreality, detachment, or being an outside observer with respect to one's thoughts, feelings, sensations, body, or actions” ([American Psychiatric Association 2013](#), p. 302). On the other hand, in the Cotard delusion (CD), mental autobiographies are acutely distorted—in such a way that patients are convinced that they are dead or that they do not exist:

An 88-year-old man with mild cognitive impairment was admitted to our hospital for treatment of a severe depressive episode. He was convinced that he was dead and felt very anxious because he was not yet buried. This delusion caused extreme suffering and made outpatient treatment impossible. ([Debruyne et al. 2009](#), p. 197)

Researchers in the field of delusion studies have debated the way in which anomalous experience leads to false belief. In this commentary I am

interested in the following questions: What cognitive architecture could, in principle, explain CD in terms of its development from depersonalization, and what exactly are the underlying differences between patients suffering from the Cotard delusion and those suffering from depersonalization disorder (DPD) but free from the Cotard delusion?

In his target paper, [Gerrans](#) explores the cognitive structure of self-awareness—the “awareness of being a unified persisting entity” ([this collection](#), p. 2). To explain the emergence of self-awareness and its loss in DPD and CD, he provides an account that integrates the appraisal theory of emotion, the simulation model of memory and prospection, and the hierarchical predictive coding model. First, based on the appraisal theory, Gerrans shows that the activation of the anterior insular cortex (AIC) allows an organism to experience the emotional significance of a relevant state by experiencing appraisal. According to [Gerrans](#), these reflexive processes are what sustain the self from moment to moment: “An organism that can use that affective information in the process is a self” ([this collection](#), p. 8). Second, the integration of affective processing and simulated episodes allows the organism to experience itself as a persisting entity overtime (see more below). Last, he endorses the predictive coding framework, according to which the human mind can be accounted for by the principle of predictive error minimization. Perception, for instance, is realized by the operation of both top-down prediction and bottom-up predictive error. If the general theoretical model is correct, it will not only apply to perception, but also to affective processing (*ibid.*, p. 9). [Gerrans](#) ([this collection](#)) applies this framework to explain the phenomenon of depersonalization and CD: Depersonalization occurs due to a failure to attribute emotional relevance to bodily states, which results from hypoactivity of the AIC. The prediction error from the mismatch between the predicted and the actual activation level of the AIC would lead to allocation of attention, the function of which, according to the predictive coding framework, is to disambiguate signals. If the prediction error cannot be cancelled and attention

cannot be diverted, increased attention brings about anxiety in DPD and CD, which is “an adaptive mechanism that primes the organism cognitively and physiologically to solve uncertainty” (*ibid.*, p. 11). This is reflected in the patients’ subjective reports concerning the loss of awareness of their bodies. This integrated theory provides an explanation of depersonalization as well as of how self-awareness is constructed through the interaction of different forms of cognitive processing.

In Gerrans’ account, the simulation system allows the organism “to *simulate* temporally distant experiences by rehearsing some of the same perceptual and emotional mechanisms activated by the simulated situation” (*ibid.*, p. 6), such that the affective associations result in integrated episodes of experience that lead to the feeling of persisting over time. I argue (1) that the simulation model should not be thought of as independent from other memory systems: without memory systems at lower levels—semantic and procedural memory systems—the simulated episodes cannot be constructed (section 2); and (2), that by considering the role of memory under the predictive coding framework, the simulation model not only plays a role in simulating temporally-distant episodes but also contributes to the knowledge required for the creation of predictive models in the present (section 3). On such a view of the simulation model, delusion can be explained and I will suggest (3) two factors contributing to the development of CD from depersonalization: the compromised decontextualized supervision system and the expectation of high precision from interoceptive signals (section 4); that is, only if these two factors are present in a depersonalized subject may CD develop.

2 The simulation model and the mental autobiography

[W]e are all virtuoso novelists, who find ourselves engaged in all sorts of behavior, more or less unified, but sometimes disunified, and we always put the best ‘faces’ on it we can. We try to make all of our material cohere into a single good story. And

that story is our autobiography. (Dennett 1992, p. 114)

As persons, our beliefs and desires are structured in a more or less coherent fashion, such that a mental autobiography—an autobiographical framework (Gerrans 2013) or narrative (Schechtman 1996)—can be attributed, which can explain our cognitive structure. Many people have proposed theoretical entities such as the “autobiographical self” (Damasio 1999, 2010), the “conceptual self” (Conway 2005; Conway et al. 2004), and the “narrative self” (Feinberg 2009), etc. to account for how one comprehends and navigates through the world and over time—that is, how one is able to make sense of external or internal signals, to have preferences, to have goals and to values, to know who oneself is, to be a diachronically persisting agent, to recall the past, and to imagine the future. In general, these different versions of the “extended self” (Gallagher 2000) are characterized by the following phenomenal and epistemic properties.

First, phenomenally, we experience ourselves as thinkers of thoughts (e.g., “I think...” or “I believe...”) and as beings who recollect the past and plan for the future; while at the same time we have a sense of ownership of relevant beliefs (e.g., “this thought is mine”). Second, subjectively, events and objects are presented in a way that manifests their relevance to the subject. In addition, epistemically, we tend to treat the self-told story as if it were highly reliable: The content is treated as objectively real, and its truthfulness is seldom questioned. This is the way we consciously comprehend the world and our place within it, and it is thought to be reliable. Accompanied by a certain degree of the “feeling of familiarity” and the “sense of pastness” (Russell 2009, p. 208), there is a degree of certainty about the veridicality of a mental autobiography. When inconsistency or non-veridicality is detected and such certainty is lost (e.g., due to introspection or contradiction to external information), the mental autobiography will be modified to re-create a new subjective reality—a new story about ourselves with more or less difference (e.g., self-deception).

Delusional patients have anomalous forms of mental autobiography: Their mental autobiographies are radically distorted, for different reasons. For instance, RZ, a 40-year-old female patient with reverse intermetamorphosis, believed that she was her father (and sometimes believed that she was her grandfather) during her assessment by Breen et al. (2000). When asked to sign her name and answer questions about her life, she signed her father's name and provided her father's personal history. She acted according to her delusional beliefs. Here we see that her mental autobiography constructs her subjective reality. Semantic dementia patients who suffer from an incapability of constructing personal futures (Irish & Piguet 2013) provide examples of the loss of partial subjective reality.² It is speculated that this form of futureless mental autobiography accounts for the higher suicide rate in semantic dementia (Hsiao et al. 2013). As we will see, patients suffering from CD also maintain a mental autobiography.³ They believe that they are dead or no longer exist: They may refuse to eat or visit the graveyard—the place in which they believe they belong. But how are mental autobiographies constructed? The rest of this section considers how memory systems and simulation models lead to the construction of a mental autobiography.

Studies on misrepresentations in memory have suggested that—against the traditional and folk-psychological idea of a “store-house” (Locke 2008), in which memory as a copy of past experience is stored for future use—memory is constructive in nature. It represents different facets of experience, which are distributed across different regions of the brain, where retrieval is realized in a process of pattern completion, which allows a subset of features to comprise a past experience (Schacter et al. 1998). The prevalence of misremembering (episodic memory in particular) and the view of con-

structive memory have led to the debate over the function of memory: If the proper function of memory is to veridically represent past experiences or events, is our memory system fundamentally defective? Or, does it serve other functions? If there is any adaptive advantage of memory systems, they must serve a function that concerns the *current and/or future* states of the organism (Westbury & Dennett 2000). New findings regarding a default-mode network suggest a “constructive episodic simulation hypothesis” (Schacter & Addis 2007a, 2007b), according to which the constructive nature of episodic memory is partially attributable to its proposed role in mentally simulating our personal futures (e.g., planning a future event). This hypothesis is supported by fMRI evidence showing that remembering the past as well as imagining the possible past and future share correlates with the activities of the default mode network (Addis et al. 2007; De Brigard et al. 2013; Szpunar et al. 2007). Therefore, it is suggested that episodic memory is adaptive in that it allows us to employ past experiences in such a way as to enable simulations of possible future episodes.

However, simulation is not realized by episodic memory alone. Though memory systems (i.e., procedural, semantic, and episodic memory) can be conceptually distinguished, they are considered parts of a “monohierarchical multimemory systems model” (Tulving 1985): Semantic memory is a specialized subsystem of procedural memory that lies at the lowest level of the hierarchy, and semantic memory in turn contains episodic memory as its specialized subsystem. The subsystems at higher levels are dependent on and supported by those at lower levels. That is, our everyday autobiographical memory is realized by multiple memory systems. For instance, a recent study has shown the importance of semantic memory in the construction of autobiographical memory: While episodic memory provides episodic details, semantic memory acts as a schema for integrating them (Irish & Piguet 2013). That is, our mental autobiographies are constructed by the interplay of multiple memory systems (e.g., Tulving's SPI model, see Tulving 1995).

² If the predictive coding framework and the role of memory for which I argued in section 3 is correct, one should expect to find an anomalous phenomenon in semantic dementia—not only with respect to one's narrative consciousness, but also with respect to one's perception.

³ It might be a contradiction in terms to claim that patients suffering from the Cotard delusion have mental autobiographies, since “auto” means “self, one's own” and “bio” means “life”. Here, it can merely be understood as a personal-level response to the system's condition.

This applies to prospection as well. Different categorizations of prospection are proposed (e.g., [Atance & O'Neill 2001](#); [Szpunar et al. 2014](#)). In this commentary, I adopt a distinction offered by [Suddendorf & Corballis \(2007\)](#), who distinguish procedural, semantic, and episodic prospection (p. 301, Figure 1). [Suddendorf & Corballis \(2007\)](#) suggest that the function of the memory and anticipatory systems is to provide behavioral flexibility; and they also examine the phylogenetic development of different memory systems. According to their model, the flexibility of anticipatory behavior supported by different memory systems can offer varies in degree. From the primitive form, procedural memory enables stimulus-driven predictions of regularities and allows behavior to be modulated by experience, such that the resulting behavior is stimulus-bound. Declarative memory provides more flexibility because it can not only be retrieved involuntarily, but can also be voluntarily triggered top-down from the frontal lobe, which enables decoupled representations that are not directly tied to the perceptual system. That is, even though we are still tied to the present in that we recall and imagine the future at the present moment, the content of representation can extend beyond the current immediate environment. Specifically, semantic memory is considered more primitive than episodic memory as it has less scope for flexibility ([Suddendorf & Corballis 2007](#)).⁴ The former, in allowing learning in one context to be voluntarily transferred to another, provides the basis for reasoning. However, this is about regularities and not particularities. Episodic memory supplements this weakness: A scenario can be simulated and pre-experienced. Through mental reconstruction or memory construction, episodic memory not only recreates past events, it also allows the learned

elements to be incorporated and arranged in a particular way in order to simulate possible futures. It thereby provides greater flexibility in novel situations and provides for the possibility of making long-term plans, extending even beyond the life-span of the individual.

To sum up, our mental autobiography is constructed through the interaction of multiple memory systems at different levels. The simulation model should not only be associated with the episodic memory system; rather, it should be understood as a hierarchical model of multiple memory systems—i.e., procedural, semantic, and episodic memory as well as procedural, semantic, and episodic prospection. In the next section I will consider the role of memory systems within the predictive coding framework.

3 Memory under the predictive coding framework

Recent development of the predictive coding framework ([Clark 2013b](#), [this collection](#); [Friston 2003](#); [Hohwy this collection](#)) provides an integrated conceptual framework for perception and action. According to the framework, the brain constantly attempts to minimize the discrepancy between sensory inputs (including exteroceptive and interoceptive signals) and the internal models of the causes of those inputs via reciprocal interactions between hierarchical levels. Each cortical level employs a generative model to predict representations of the subordinate level, to which the prediction is sent via top-down projections—the bottom-up signal is the prediction error. Prediction error minimization can be achieved in a number of ways ([Clark this collection](#); [Hohwy this collection](#)); but in general, errors can be minimized either by updating generative models to fit the input or by carrying out actions to change the world to fit the model. In the target paper, [Gerrans](#) integrates appraisal processing into the predictive coding framework; however, he treats only the simulation model as a mechanism for simulating temporally distant experiences ([this collection](#), pp. 6–8). In this section, I propose that under the predictive coding framework, the simulation model serves the function of updating

⁴ [Tulving \(2005\)](#) and [Suddendorf & Corballis \(2007\)](#) argue that episodic memory emerges later in the course of evolution and belongs uniquely to human beings. Even if there is evidence suggesting the existence of episodic-like memory—memory encoding “what”, “where”, and “when” information—in non-human creatures (e.g., Western scrub jays; [Clayton 2003](#); [Clayton & Dickinson 1998](#)), [Tulving \(2005\)](#) argues that these phenomena can be explained merely by semantic memory. In a recent paper, [Corballis \(2013\)](#) changes the claim he makes in the earlier article ([Suddendorf & Corballis 2007](#)) and argues that mental time travel also exists in rats, and that the difference between this and human mental time travel is simply the degree of complexity.

the knowledge required for successful prediction, which constitutes perception and affective experience.

How can we understand the role of memory or the simulation system under the predictive coding framework?⁵ Here I examine how memory systems can be incorporated into the framework. According to the predictive coding framework, perceiving is distinct from the traditional model of perception; instead, it is:

to use whatever stored knowledge is available to guide a set of guesses about [...external causes], and then to compare those guesses to the incoming signal, using residual errors to decide between competing guesses and (where necessary) to reject one set of guesses and replace it with another. (Clark 2013a, p. 743)

That is, perception is knowledge-driven and top-down, rather than stimulus-driven and bottom-up. “Stored knowledge” refers to a repertoire of prior beliefs or knowledge—the belief of the likelihood of a hypothesis or guess irrespective of sensory input. It is acquired or shaped by learning from past experience—or, in other words, it is a modification of parameters in order to minimize prediction error.⁶

Moshe Bar (2009) suggests that “our perception of the environment relies on memory as much as it does on incoming information” (p. 1235). Since we seldom encounter completely novel objects or events, our systems rely on representations stored in memory systems to generate predictions. According to Bar’s “analogy-association-prediction” framework (Bar & Neta 2008), once there is a sensory input, the brain actively generates top-down guesses in order to figure out what that input looks like (analogy); the match triggers activation of associated rep-

resentations (association), which allows predictions of what is likely to happen in the relevant context and environment (prediction). Thus, instead of aiming to answer the question “what is this?”, perception studies should answer the question “what is this *like*?” or “what does this resemble?”: Brains proactively compare incoming signals with existing information gained in the past (see Bar 2009, Figure 1 & Figure 2). Bar (2009) suggests that predictions also influence memory encoding. Memory systems primarily encode that which differs from memory-based prediction, and if sensory information meets the prediction, the information is less likely to encode (Bar 2009, p. 1240).

This account provides a new view of the concepts of encoding, retrieval, and reconsolidation. The older view describes encoding as the process by which incoming information is stored for later retrieval, and retrieval as a process involved in utilizing encoded information in reviving past events. Nevertheless, under the predictive coding framework, when discrepancy between prediction and perceptual information occurs, encoding is the process of minimizing prediction error—the adaptation of the model to reduce discrepancy based on the forward-feeding, bottom-up input from its subordinate level. Retrieval is then regarded as the process of utilizing this knowledge for predictive model construction.

Accordingly, I suggest that the role of memory systems is to update the knowledge required for successful predictions of the organism’s current (and future) informational state. That is, under the predictive coding framework, our perception is knowledge-driven, and knowledge is experience-based. The mechanisms of our memory systems allow the knowledge required for the construction of predictive models to be updated based on experience. Prediction error can trigger encoding that modifies our knowledge, which then optimizes the predictive model to achieve prediction error minimization. In addition, as we will see later in this section, the development of episodic memory and mind-wandering allows us to generate new knowledge.

This knowledge-driven perception is realized by a multi-layer hierarchical structure in

⁵ Felipe De Brigard (2012) considers how the predictive coding framework can predict remembering. He modifies Anderson’s Adaptive Control of Thought-Rational model (Anderson & Schooler 2000); here the probability of a memory retrieval can be calculated based on how well memory retrieval will minimize prediction error given the cost of the retrieval and the current context. Here, however, I shall not consider the retrieval of individual memories; instead I focus on the role of the memory systems within the framework.

⁶ See Clark (2013a) for the problem of the acquisition of the very first prior knowledge.

which “each layer is trying to build knowledge structures that will enable it to generate the patterns of activity occurring at the level below” (Clark 2013a, p. 483). The information encoded at each level is distinct: At higher hierarchical levels, the representations become more abstract and involve a larger spatial and temporal dimension: The predictive models generated not only represent the immediate state of the system or environment but also the system in relation to the spatially and temporally-extended environment. Moreover, the higher-level knowledge also supports predicting how sensory signals will change and evolve over time. It allows one to predict the future and execute long-term plans involving multiple steps. The hierarchical structure is crucial to our capacity to comprehend the world, which is highly structured, with regularity and patterns at multiple spatial and temporal scales and interacting and complexly-nested causes (Clark 2013a).

I suggest that each level of knowledge has an updating mechanism, which is consistent with Tulving’s (1985) monohierarchical multimemory systems model and Suddendorf & Corballis’ (2007) model of memory and prospection. Procedural memory at the lowest level is involved in the sensori-motor predictive function: It updates the procedural knowledge required for predicting the states in which given actions are executed. Whereas implicit memory is mainly involved in immediate responses to current stimuli, declarative or explicit memory (episodic memory in particular) contributes to the construction of a model of the system itself and its environment with spatial and temporal dimensions. It supplements higher-level knowledge structures for the construction of a generative model, which explains actual states and predicts possible changes and actions for reaching desired states. Under the predictive coding framework, the semantic memory system, which allows learning in one context to be transferred to another, supports semantic knowledge, which in turn provides regularities in the construction of predictive models (e.g., during reading). And episodic memory, together with semantic memory, supports the knowledge required to construct a model of one’s autobiography—a

model of one’s own relevant past and potential future. However, it is worth noting that our mental autobiography is not realized by knowledge at a single hierarchical level; instead, it is constructed through the interplay of the mechanisms at multiple levels.

In addition to its contribution to an autobiographical-scale model, episodic memory, along with other memory systems, also generates new knowledge by simulation. Bar (2007) proposes that:

[the] primary role [of mental time travel] is to create new ‘memories’. We simulate, plan and combine past and future in our thoughts, and the result might be ‘written’ in memory for future use. These simulated memories are different from real memories in that they have not happened in reality, but both real and simulated memories could be helpful later in the future by providing approximated scripts for thought and action. (p. 286)

This is supported by the evidence that mind-wandering—that is, having thoughts that are unrelated to the current demands of the external environment (Schooler et al. 2011)—is beneficial to autobiographical planning and creative problem solving (Mooneyham & Schooler 2013).⁷

The role of memory systems under the hierarchical predictive coding framework is consistent with the function of memory and the concept of a memory system proposed by De Brigard (2013). Following Carl F. Craver’s idea of a mechanistic role function (2001), De Brigard argues for a larger cognitive system of “episodic hypothetical thinking”, which includes

⁷ This is related to the philosophical debate on whether one can gain new knowledge from imagination or a purely mental activity, as was famously denied by Sartre (1972) and Wittgenstein (1980) (for a general discussion, see Stock 2007). It is worth noting that if the predictive coding framework is correct, the concept of “knowledge” may be revised: Knowledge may depart from veridicality; instead, it is close to information that can provide successful predictions. Thus, under the predictive coding framework, the only kind of knowledge Sartre recognizes (as cited in Stock 2007, p. 176)—observational knowledge—is not substantially different from other kinds of knowledge, because the knowledge gained through perception cannot be conceptually distinguished from those that are not: Gaining knowledge at each level is all about optimizing the predictions of lower levels.

future simulation and past counterfactual simulations: To determine the mechanistic function of memory we require an investigation into the way that its components contribute to the system, and then of how memory contributes to the functioning of the organism, helping it to reach goals at higher levels. It is worth noting that these concepts of memory function and malfunction are different to traditional ones: The distinction between memory function and malfunction is not equivalent to the distinction between remembering and misremembering or veridical representation and misrepresentation. Under the predictive coding framework, memory function can be regarded as updating knowledge for predictive model construction. Likewise, memory function and malfunction are independent from the generation of a predictive model that succeeds or fails in representing the world. That is, certain misrepresentations can lead to error minimization; furthermore, it is possible for misrepresentation rather than veridical representation to lead to a generative model.

4 From depersonalization to Cotard delusion

If the predictive coding framework is correct, it provides a new view not only on memory function but also on how we think about memory systems and the relation between memory and other cognitive systems. This framework provides a theory about the role of simulation models in the relationship between reflexive forms of self-consciousness and the narrative self (Hohwy 2007). It provides a theoretical explanation of the finding that memory systems are also involved in perception⁸ and interoception. This implies that we not only simulate offline (e.g., mental time travel, mind-wandering), but also simulate online. The simulated model provides us with a subjective reality through which we see the external world and ourselves. It is transparent and immediate: We experience it as objectively real and we directly interact with what is represented.

⁸ This is consistent with the evidence that memory influences perception (e.g., Summerfield et al. 2011).

However, this characteristic is absent in patients suffering from depersonalization. Depersonalization is an example of how one can become detached from one's simulated model of oneself: One's mental autobiography is no longer direct, and one experiences a sense of distance from the model.⁹ Gerrans (this collection) suggests that the loss of sense of presence in depersonalized patients results from a failure to minimize prediction error from the hypoactivity of the AIC—the activation of which informs us of the significance of external or internal information. Gerrans' theory is based on Seth et al.'s idea of interoceptive inference (or interoceptive predictive coding; see also Seth this collection), according to which predictive coding not only applies to exteroception but also to interoception, and emotional states, including the sense of presence, arise from interoceptive prediction successfully matched to actual interoceptive signals (Seth 2013; Seth et al. 2011). As it is suggested that the AIC is suggested to be the correlate of the integration of exteroceptive and interoceptive signals and that it plays a role in maintaining a salience network for the relevant states, the hypoactivity of the AIC leads to the failure to associate affective significance with bodily states. As Gerrans suggests, “not all higher level control systems can and do smoothly cancel prediction errors generated at lower levels” (this collection, p. 9). Because the coding formats at each level are distinct, the coding format of low-level processing is opaque to introspection (p. 9). The problems faced by depersonalized patients can be accounted for by the prediction error based on persisting, unexpected hypoactivity. Attention is then directed towards resolving the prediction error. Gerrans' proposal is that an inability to explain away the surprisal and this increased attention causes anxiety in DPD. Here, CD can be seen as a strategy for some systems to react to anxiety in order to minimize the prediction error.

As Gerrans suggests, “[d]elusions are best conceptualized as higher-level responses to pre-

⁹ In contrast to depersonalization, derealization refers to the “[e]xperiences of unreality of detachment with respect to surroundings” (American Psychiatric Association 2013, p. 302)—patients suffer from detachment from the simulated model of the environment.

diction error which, however, cannot cancel those errors” ([this collection](#), p. 10). That is, even though not all prediction error can be successfully cancelled, the brain—the organ that constantly minimizes prediction error, according to predictive coding framework—still tries to modify its model in order to decrease surprisal, though unsuccessfully. If what I have suggested in the last section is correct, the function of memory systems is to update knowledge contributing to the construction of predictive models in order to minimize prediction error. The anomalous model of CD is thus one constructed by the hierarchical simulation model to match the hypoactivity of the AIC—the loss of appraisal that represents the significance of self-related information. To construct a model in which oneself is dead or does not exist cannot successfully explain away the prediction error—since one still has the experience of a bodily state—it may nevertheless be the best solution the given system can come up with in order to cope with the increased anxiety resulting from increased attention.

However, this still leaves us with the question of why some depersonalized patients develop CD, whereas most of them do not develop this delusion. [Gerrans \(2014\)](#) suggests that the difference between delusional and non-delusional minds lies in differences in the default mode network, which include information that triggers activity, hyperactivity, and hyperconnectivity, interaction with the salience system, and absent or impaired “decontextualized supervision” (pp. 73–74). Decontextualized supervision allows one to “reason about *oneself* using impersonal, objective rules of inference” (p. 76).¹⁰ The activity of its circuit is anti-correlated with the activity of the default mode network (pp. 83–84) because of the limited cognitive resources for high-level metacognitive processes. Gerrans suggests that delusional thoughts arise from the system’s failure to balance this allocation; thus they slip through the supervision system.

Nevertheless, the existence of decontextualized supervision explains how anomalous forms of predictive models—which would be suppressed in non-delusional subjects—could emerge, but it does not account for the model’s relation to anomalous experience or to the way in which the content of delusion is constructed (e.g., Cotard delusion). I therefore propose that a delusional mind does not only result from a compromised decontextualized supervision; it also results from an aberrant precision expectation¹¹ of exteroceptive or interoceptive signals. [Jakob Hohwy \(2013\)](#) proposes the notion of uncertainty expectations: We predict the causal structure of the world (and of one’s own bodily state), as well as the level of uncertainty in the environment, which allows us to respond to the external environment under various levels of uncertainty. The strength of prediction error is proportional to the expected certainty: When the uncertainty level is expected to be higher (due to external or internal noise), the prior model is weighted higher, whereas expected low uncertainty gives more weight to bottom-up prediction error. According to [Hohwy \(2013\)](#), delusion arises when precision expectation is either too high or too low, and those in between would report only the anomalous experience, without forming a delusion. In the case of Cotard delusion developed from depersonalization, when one has the expectation of high precision, the system tends to be driven by the bottom-up predictive error of unexpected hypoactivity of the AIC, rather than the prior model. One is, therefore, more likely to revise the model in order to explain away the surprisal resulting from the mismatch between the actual and predicted activation level of the AIC; that is, the systems of patients suffering from CD are driven by an urge to modify their top-down predictive models in order to conform to the loss of AIC activity. The construction of the model in CD is considered an attempt to minimize prediction error.

Finally, explaining delusion under the predictive coding framework provides new understanding to the debate between one- and two-

¹⁰ The system of decontextualized supervision is distinct from the semantic memory system discussed in the last section: The latter provides objective elements for the construction of a contextualized autobiographical episode, while the former supervises autobiographical episodes by utilizing decontextualized reasoning.

¹¹ “Precision” is also used to refer to the precision of inferences about hidden causal structures (e.g., in [Friston et al. 2013](#)). Here and in [Hohwy \(2013\)](#) it indicates the precision of incoming signals.

stage models of delusion. The one-stage model holds that anomalous experience only suffices to explain the occurrence of delusion (Gerrans 2002; Maher 1974, 1988); according to two-stage model, however, other cognitive disruption is required to explain the content of the delusion in particular (Young & De Pauw 2002). However, if the predictive coding framework is correct, the clear distinction between experience and rationalization assumed in the traditional discussion does not exist: Perception, cognition, and action are now considered continuous and highly integrated (Clark 2013b; Hohwy & Rajan 2012). Experience and rationalization are different layers of abstraction within the very same process of prediction error minimization under the predictive coding framework.

5 Conclusion

In his target paper, Philip Gerrans proposes a theory of self-awareness that integrates the predictive coding framework, the appraisal theory, and the simulation model. It accounts for the loss of self-awareness in DPD and CD, and provides a new understanding of patients' anxiety. In this commentary, I have proposed (1) that the simulation model should be considered a hierarchical model involving multiple memory systems—namely, it is constituted by procedural, semantic, and episodic memory and prospective (section 2); and (2) that the function of memory systems or simulation models, under the predictive coding framework, is to update the knowledge required for successful prediction (section 3). This implies that memory function and malfunction are independent from the generation of a predictive model that succeeds or fails in representing the world, since it is possible that misrepresentation rather than veridical representation leads to a generative model that minimizes prediction error. Based on such view of the simulation model, CD can be regarded as the modification of top-down prediction in an attempt to explain away the prediction error resulting from unexpected hypoactivity of the AIC. I also suggested (3) that a combination of two factors is necessary for the occurrence of CD from depersonalization: the

compromised decontextualized supervision system and the expectation of high precision of interoceptive signals (section 4).

If both the general framework and my suggestions are correct, there are a number of issues worthy of further investigation: First, if the model that explains the symptoms of CD is created by the system in order to minimize prediction error from hypoactivity of the AIC, with the aim of affording relief from anxiety, it is expected that the change of prediction may be accompanied by minimized prediction error or/and prediction error from other unpredicted activities. In the case of Cotard delusion, the new model—the model of the organism's death or non-existence—would encounter new kinds of prediction error due to information about bodily states, instead of a lack of emotional significance. This may as well be the kind of prediction error that cannot be cancelled top-down and which can be expected to lead to anxiety based on Gerrans' theory. Therefore, the anxiety characteristic of the Cotard delusion is speculated to be the result of different prediction errors from patients suffering from Cotard syndrome. Studies on the difference between the anxiety present in DPD and that in CD can support or refutation of the framework proposed. Furthermore, it is worth noting that not all patients with the CD suffer from anxiety. For example, in Berrios & Luque's (1995) analysis of 100 cases, anxiety is reported in only 65% of subjects, and patients were categorized: Cotard type I patients showed no affective component, whereas type II patients showed depression and anxiety. Can the proposed framework account for both types of patients?

Another interesting question for future research is whether we can better understand the relation between the simulation model and affective processing within the predictive coding framework, and whether an explanation of this would be consistent with the existing evidence relating to emotional memory (e.g., LaBar & Cabeza 2006). Affective processing can influence encoding and retrieval of memories, whereas simulating possible episodes is thought to help rehearse affective responses. One possible avenue might be the investigation of the influence

of different forms of simulation on affective processing (e.g., memory retrieval from a field or an observer perspective; Berntsen & Rubin 2006), and further on one's awareness of one's future and past (Wilson & Ross 2003): How can this be accounted for by the principle of prediction error minimization? Does the simulation of potential affective responses optimize prediction and reduce potential error in the future? The simulation and integration of future potential changes into the model of one's autobiography is thought to potentially contribute to the prevention of dramatic changes in one's model at higher levels, and to maintain mental autobiographies that are more consistent across time.

Acknowledgments

I am grateful to Thomas Metzinger and Jennifer M. Windt, as well as two reviewers, for their critical and constructive comments.

References

- Addis, D. R., Wong, A. T. & Schacter, D. L. (2007). Remembering the past and imagining the future: Common and distinct neural substrates during event construction and elaboration. *Neuropsychologia*, 45 (7), 1363-1377. [10.1016/j.neuropsychologia.2006.10.016](https://doi.org/10.1016/j.neuropsychologia.2006.10.016)
- American Psychiatric Association, (2013). *The diagnostic and statistical manual of mental disorders*. Arlington, VA: American Psychiatric Publishing.
- Anderson, J. R. & Schooler, L. J. (2000). The adaptive nature of memory. In E. Tulving & F. I. M. Craik (Eds.) *The Oxford handbook of memory* (pp. 557-570). New York, NY: Oxford University Press.
- Atance, C. M. & O'Neill, D. K. (2001). Episodic future thinking. *Trends in Cognitive Sciences*, 5 (12), 533-539. [10.1016/s1364-6613\(00\)01804-0](https://doi.org/10.1016/s1364-6613(00)01804-0)
- Bar, M. (2007). The proactive brain: Using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, 11 (7), 280-289. [10.1016/j.tics.2007.05.005](https://doi.org/10.1016/j.tics.2007.05.005)
- (2009). The proactive brain: Memory for predictions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364 (1521), 1235-1243. [10.1098/rstb.2008.0310](https://doi.org/10.1098/rstb.2008.0310)
- Bar, M. & Neta, M. (2008). The proactive brain: Using rudimentary information to make predictive judgments. *Journal of Consumer Behaviour*, 7 (4-5), 319-330. [10.1002/cb.254](https://doi.org/10.1002/cb.254)
- Berntsen, D. & Rubin, D. C. (2006). Emotion and vantage point in autobiographical. *Cognition and Emotion*, 20 (8), 1193-1215. [10.1080/02699930500371190](https://doi.org/10.1080/02699930500371190)
- Berrios, G. E. & Luque, R. (1995). Cotard's syndrome: Analysis of 100 cases. *Acta Psychiatrica Scandinavica*, 91 (3), 185-188. [10.1111/j.1600-0447.1995.tb09764.x](https://doi.org/10.1111/j.1600-0447.1995.tb09764.x)
- Breen, N., Caine, D., Coltheart, M., Hendy, J. & Roberts, C. (2000). Towards an understanding of delusions of misidentification: Four case studies. *Mind & Language*, 15 (1), 74-110. [10.1111/1468-0017.00124](https://doi.org/10.1111/1468-0017.00124)
- Clark, A. (2013a). Expecting the world: Perception, prediction, and the origins of human knowledge. *Journal of Philosophy*, 110 (9), 469-496.
- (2013b). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36 (3), 181-204. [10.1017/S0140525X12000477](https://doi.org/10.1017/S0140525X12000477)
- (2015). Embodied prediction. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Clayton, N. S. & Dickinson, A. (1998). Episodic-like memory during cache recovery by scrub jays. *Nature*, 395 (6699), 272-274. [10.1038/26216](https://doi.org/10.1038/26216)

- Clayton, N. S., Bussey, T. J. & Dickinson, A. (2003). Can animals recall the past and plan for the future? *Nature Reviews Neuroscience*, 4 (8), 685-691. [10.1038/nrn1180](https://doi.org/10.1038/nrn1180)
- Conway, M. A. (2005). Memory and the self. *Journal of memory and language*, 53 (4), 594-628. [10.1016/j.jml.2005.08.005](https://doi.org/10.1016/j.jml.2005.08.005)
- Conway, M. A., Meares, K. & Standart, S. (2004). Images and goals. *Memory*, 12 (4), 525-531. [10.1080/09658210444000151](https://doi.org/10.1080/09658210444000151)
- Corballis, M. C. (2013). Mental time travel: A case for evolutionary continuity. *Trends in Cognitive Sciences*, 17 (1), 5-6. [10.1016/j.tics.2012.10.009](https://doi.org/10.1016/j.tics.2012.10.009)
- Craver, C. F. (2001). Role functions, mechanisms, and hierarchy. *Philosophy of Science*, 68 (1), 53-74. [10.1086/392866](https://doi.org/10.1086/392866)
- Critchley, H. D. (2005). Neural mechanisms of autonomic, affective, and cognitive integration. *Journal of Comparative Neurology*, 493 (1), 154-166. [10.1002/cne.20749](https://doi.org/10.1002/cne.20749)
- Damasio, A. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. New York, NY: Harcourt Brace.
- (2010). *Self comes to mind: Constructing the conscious brain*. New York, NY: Pantheon.
- De Brigard, F. (2012). Predictive memory and the surprising gap. *Frontiers in Psychology*, 3 (420). [10.3389/fpsyg.2012.00420](https://doi.org/10.3389/fpsyg.2012.00420)
- (2013). Is memory for remembering? Recollection as a form of episodic hypothetical thinking. *Synthese*, 191 (2), 1-31. [10.1007/s11229-013-0247-7](https://doi.org/10.1007/s11229-013-0247-7)
- De Brigard, F., Addis, D. R., Ford, J. H., Schacter, D. L. & Giovanello, K. S. (2013). Remembering what could have happened: Neural correlates of episodic counterfactual thinking. *Neuropsychologia*, 51 (12), 2401-2414. [10.1016/j.neuropsychologia.2013.01.015](https://doi.org/10.1016/j.neuropsychologia.2013.01.015)
- Debruyne, H., Portzky, M., Van den Eynde, F. & Aude-naert, K. (2009). Cotard's syndrome: A review. *Current psychiatry reports*, 11 (3), 197-202. [10.1007/s11920-009-0031-z](https://doi.org/10.1007/s11920-009-0031-z)
- Dennett, D. C. (1992). The self as a center of narrative gravity. In F. S. Kessel, P. M. Cole & D. L. Johnson (Eds.) *Self and consciousness: Multiple perspectives* (pp. 103-115). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Feinberg, T. E. (2009). *From axons to identity: Neurological explorations of the nature of the self*. New York, NY: WW Norton & Company.
- Friston, K. (2003). Learning and inference in the brain. *Neural Networks*, 16 (9), 1325-1352. [10.1016/j.neunet.2003.06.005](https://doi.org/10.1016/j.neunet.2003.06.005)
- Friston, K., Schwartenbeck, P., Fitzgerald, T., Moutoussis, M., Behrens, T. & Dolan, R. J. (2013). The anatomy of choice: Active inference and agency. [Hypothesis & Theory]. *Frontiers in Human Neuroscience*, 7 (598). [10.3389/fnhum.2013.00598](https://doi.org/10.3389/fnhum.2013.00598)
- Gallagher, S. (2000). Philosophical conceptions of the self: Implications for cognitive science. *Trends in Cognitive Sciences*, 4 (1), 14-21. [10.1016/S1364-6613\(99\)01417-5](https://doi.org/10.1016/S1364-6613(99)01417-5)
- Gerrans, P. (2002). A one-stage explanation of the Cotard delusion. *Philosophy, Psychiatry, & Psychology*, 9 (1), 47-53. [10.1353/ppp.2003.0007](https://doi.org/10.1353/ppp.2003.0007)
- (2013). Delusional attitudes and default thinking. *Mind & Language*, 28 (1), 83-102. [10.1111/mila.12010](https://doi.org/10.1111/mila.12010)
- (2014). *Measure of madness: Philosophy of mind, cognitive neuroscience, and delusional thought*. Cambridge, MA: MIT Press.
- (2015). All the self we need. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Hohwy, J. (2007). The sense of self in the phenomenology of agency and perception. *Psyche*, 13 (1), 1-20.
- (2013). Delusions, illusions and inference under uncertainty. *Mind & Language*, 28 (1), 57-71. [10.1111/mila.12008](https://doi.org/10.1111/mila.12008)
- (2015). The neural organ explains the mind. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Hohwy, J. & Rajan, V. (2012). Delusions as forensically disturbing perceptual inferences. *Neuroethics*, 5 (1), 5-11. [10.1007/s12152-011-9124-6](https://doi.org/10.1007/s12152-011-9124-6)
- Hsiao, J. J., Kaiser, N., Fong, S. & Mendez, M. F. (2013). Suicidal behavior and loss of the future self in semantic dementia. *Cognitive and behavioral neurology: official journal of the Society for Behavioral and Cognitive Neurology*, 26 (2), 85-92. [10.1097/WNN.0b013e31829c671d](https://doi.org/10.1097/WNN.0b013e31829c671d)
- Irish, M. & Piguet, O. (2013). The pivotal role of semantic memory in remembering the past and imagining the future. *Frontiers in Behavioral Neuroscience*, 7 (27). [10.3389/fnbeh.2013.00027](https://doi.org/10.3389/fnbeh.2013.00027)
- LaBar, K. S. & Cabeza, R. (2006). Cognitive neuroscience of emotional memory. *Nature Reviews Neuroscience*, 7 (1), 54-64. [10.1038/nrn1825](https://doi.org/10.1038/nrn1825)
- Locke, J. (2008). *An essay concerning human understanding*. Oxford, UK: Oxford University Press.
- Maher, B. A. (1974). Delusional thinking and perceptual disorder. *Journal of Individual Psychology*, 30 (1), 98-113.
- (1988). Anomalous experience and delusional thinking: The logic of explanations. In T. F. Oltmanns & B. A. Maher (Eds.) *Delusional beliefs* (pp. 15-33). Oxford, UK: John Wiley & Sons.

- Mooneyham, B. W. & Schooler, J. W. (2013). The costs and benefits of mind-wandering: A review. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 67 (1), 11-18. [10.1037/a0031569](#)
- Russell, B. (2009). *The analysis of mind*. Auckland, NZ: The Floating Press.
- Sartre, J.-P. (1972). *The psychology of imagination*. Oxford, UK: Blackwell.
- Schacter, D. L. & Addis, D. R. (2007a). The cognitive neuroscience of constructive memory: Remembering the past and imagining the future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362 (1481), 773-786. [10.1098/rstb.2007.2087](#)
- (2007b). Constructive memory: The ghosts of past and future. *Nature*, 445 (7123), 27-27. [10.1038/445027a](#)
- Schacter, D. L., Norman, K. A. & Koutstaal, W. (1998). The cognitive neuroscience of constructive memory. *Annual Review of Psychology*, 49 (1), 289-318. [10.1146/annurev.psych.49.1.289](#)
- Schechtman, M. (1996). *The constitution of selves*. New York, NY: Cornell University Press.
- Schooler, J. W., Smallwood, J., Christoff, K., Handy, T. C., Reichle, E. D. & Sayette, M. A. (2011). Meta-awareness, perceptual decoupling and the wandering mind. *Trends in Cognitive Sciences*, 15 (7), 319-326. [10.1016/j.tics.2011.05.006](#)
- Ségla, J. (1897). *Le délire des négations: séméiologie et diagnostic*. Paris, FR: Masson, Gauthier-Villars.
- Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, 17 (11), 565-573. [10.1016/j.tics.2013.09.007](#)
- (2015). The cybernetic Bayesian brain. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Seth, A. K., Suzuki, K. & Critchley, H. D. (2011). An interoceptive predictive coding model of conscious presence. *Frontiers in Psychology*, 2 (395). [10.3389/fpsyg.2011.00395](#)
- Shorvon, H. (1946). The depersonalization syndrome. *Proceedings of the Royal Society of Medicine*, 39 (12), 779-792.
- Steinberg, M. (1995). *Handbook for the assessment of dissociation: A clinical guide*. Washington, DC: American Psychiatric Press.
- Stock, K. (2007). Sartre, Wittgenstein and learning from imagination. In P. Goldie & E. Schellekens (Eds.) *Philosophy and conceptual art* (pp. 171-194). Oxford, UK: Oxford University Press.
- Suddendorf, T. & Corballis, M. C. (2007). The evolution of foresight: What is mental time travel, and is it unique to humans? *Behavioral and Brain Sciences*, 30 (3), 299-313. [10.1017/S0140525X07001975](#)
- Summerfield, J. J., Rao, A., Garside, N. & Nobre, A. C. (2011). Biasing perception by spatial long-term memory. *The Journal of Neuroscience*, 31 (42), 14952-14960. [10.1523/jneurosci.5541-10.2011](#)
- Szpunar, K. K., Watson, J. M. & McDermott, K. B. (2007). Neural substrates of envisioning the future. *Proceedings of the National Academy of Sciences*, 104 (2), 642-647. [10.1073/pnas.0610082104](#)
- Szpunar, K. K., Spreng, R. N. & Schacter, D. L. (2014). A taxonomy of prospection: Introducing an organizational framework for future-oriented cognition. *Proceedings of the National Academy of Sciences*
- Tulving, E. (1985). How many memory systems are there? *American Psychologist*, 40 (4), 385-398. [10.1037/0003-066X.40.4.385](#)
- (1995). Organization of memory: Quo vadis. *The Cognitive Neurosciences*, 839-847.
- (2005). Episodic memory and autonoesis: Uniquely human. In H. S. Terrace & J. Metcalfe (Eds.) *The missing link in cognition: Origins of self-reflective consciousness* (pp. 3-56). Oxford, UK: Oxford University Press.
- Westbury, C. & Dennett, D. C. (2000). Mining the past to construct the future: Memory and belief as forms of knowledge. In D. L. Schacter & E. Scarry (Eds.) *Memory, brain, and belief* (pp. 11-32). Cambridge, MA: Harvard University Press.
- Wilson, A. & Ross, M. (2003). The identity function of autobiographical memory: Time is on our side. *Memory*, 11 (2), 137-149. [10.1080/741938210](#)
- Wittgenstein, L. (1980). *Remarks on the philosophy of psychology*. Oxford, UK: Blackwell.
- Young, A. W. & De Pauw, K. W. (2002). One stage is not enough. *Philosophy, Psychiatry, & Psychology*, 9 (1), 55-59. [10.1353/ppp.2003.0019](#)

Metamisery and Bodily Inexistence

A Reply to Ying-Tung Lin

Philip Gerrans

The difference between the Cotard Depersonalisation and Depersonalisation Disorder may consist, not only in the fact that the Cotard delusion is a response to prediction error affective/bodily information, but the level in the predictive processing hierarchy at which predictions about bodily information are violated.

Keywords

Cotard delusion | Depersonalisation disorder | Interoception | Predictive coding | Self awareness

Author

[Philip Gerrans](#)

philip.gerrans@adelaide.edu.au
University of Adelaide
Adelaide, Australia

Commentator

[Ying-Tung Lin](#)

linyingtung@gmail.com
國立陽明大學
National Yang-Ming University
Taipei, Taiwan

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Prediction error and veridicality

My explanation of Depersonalisation Disorder (DPD) argued that the characteristic experience is shared by people who suffer from the Cotard Delusion (CD). The difference between the two conditions is that the person with DPD does not develop a delusional response to her experience of de-affectualisation. She simply reports as it is: “I *feel as if* my experiences do not belong to me”. The person with Cotard, however develops an explanation of that feeling and identifies with it “I no longer exist”. In commenting on this proposal Ying-Tung Lin opens up a range of new possibilities for cognitive the-

orizing. The first is that the predictive coding approach provides a new framework for cognitive theorizing which improves on “second factor” approaches to delusion. The second is that attention to the predictive nature of the processes which generate experience might suggest an important difference between the two conditions: namely the role of the Anterior Insular Cortex (AIC).

One way to approach the phenomenon would be to ask why the person with DPD seems to be able to understand that her experience is not veridical while the person with CD

does not (*modulo* all the *caveats* about the epistemic status of delusional attitudes). The CD patient for example does not say “It feels as if I don’t exist” she says “I don’t exist”. This way of approaching the problem fits with a now standard approach to delusion, that argues that there are (at least) two stages of cognitive processing involved in delusion formation. The first generates an anomalous experience and the second generates a delusional response to that experience.

Ying-Tung Lin however, following Hohwy and Clark, explains delusion in terms of the attempt by higher order control systems to account for surprisal in a predictive coding hierarchy. The radical aspect of these ideas is that neither the precipitating experience nor the delusional response need be conceived of as the result of cognitive malfunction. Because there is no intrinsic connection between error minimization and malfunction “certain misrepresentations can lead to error minimization; furthermore, it is possible for misrepresentation rather than veridical representation to lead to a generative model” (Lin this collection, p. 8).

Ying-Tung Lin’s commentary applies these ideas to the Cotard delusion, arguing that it is a model that minimises the prediction error represented by depersonalisation experience. Her target is to describe a

cognitive architecture [that] could, in principle, explain CD in terms of its development from depersonalization, and what exactly are the underlying differences between patients suffering from the Cotard delusion and those suffering from depersonalization disorder (DPD) but free from the Cotard delusion? (Lin this collection, p. 2)

2 The sense of presence

Before I make some comments, I want to highlight the original aspects of her account and show how it can explain how experience acquires a quality of “mineness” or “sense of presence”, that is of belonging to a self. We can then use the predictive coding framework to ex-

plain how the sense of presence can go missing. Loss of the sense of presence signals a prediction error which then requires a higher-level system to build a predictive model that fits that error.

The first point to note is that on the most radical interpretation of predictive coding ideas the veridicality of representation is a corollary of cognition not its primary goal. The primary goal of a cognitive system is to predict its own informational states consequent on its actions (broadly construed to include internal regulatory actions). The point is not just that the objects of experience are constructed and hence may be illusory or misrepresented. Rather veridicality of experience is secondary to the accuracy with which cognitive process predicts the flow of information in sensory systems. As she says in the case of perception this means that “instead of aiming to answer the question ‘what is this?’ perception studies should answer the question ‘... what does this resemble?’” (Lin this collection, p. 6). This formulation captures the idea that the visual system, for example, is not passively registering retinal information and constructing a representation of the external world, but using a model which predicts the flow of information coming from the retina.

The first step is to apply the same idea to interoception. We see that the mind is not passively registering changes in body state and constructing a model of the body accordingly but predicting the flow of bodily information in cognitive context. Those contexts range from maintenance of homeostasis to the use of affective experience to inform decision-making and reflective cognition. Thus when I think about the past or future these episodes of retrospection or prospection are infused with affective significance.

The radical import for the understanding of pathologies of self-representation is very elegantly brought out by her discussion. Ying-Tung Lin in effect argues that the experience of the self in autobiographical episodes is no more direct than experience of the world in perception or of past events in memory. In each case no object is directly represented or experienced. Rather the relevant object in each case (object

of perception, remembered event, or self in the case of first person awareness) is *inferred* as a part of a process of optimizing predictive accuracy in specific cognitive contexts.

As many have argued the role of the Anterior Insular Cortex (AIC) is to integrate and represent affective information: i.e., those bodily states, which tell the organism how it is faring in the world, actual, imagine or remembered. The point to recall from Ying-Tung Lin's account is that the AIC is not representing a self but constructing and optimizing a model that predicts the flow of affectively-charged bodily information.

This is why when AIC is hypoactive the subject feels a loss of subjective presence, reported as depersonalization. In particular the patient has a loss of subjective presence for her own body: she registers changes in body state but they do not feel affectively significant for her. Because that lack of feeling is not predicted she then reports it in the vocabulary of DPD.

Why does the DPD patient not proceed to something like the Cotard delusion? According to Ying-Tung Lin whether a delusion is formed depends on the degree of precision assigned to the information produced by hypoactivity in the AIC.

In the case of Cotard delusion developed from depersonalization, when one has the expectation of high precision, the system tends to be driven by the bottom-up predictive error of unexpected hypoactivity of the AIC, rather than the prior model. One is, therefore, more likely to revise the model in order to explain away the surprisal resulting from the mismatch between the actual and predicted activation level of the AIC; that is, the systems of patients suffering from CD are driven by an urge to modify their top-down predictive models in order to conform to the loss of AIC activity. The construction of the model in CD is considered an attempt to minimize prediction error.

3 Conclusion

Reading over this account I wonder if there is an alternative interpretation available consistent with the predictive coding account. It is consist-

ent with the view that patterns of activity in the AIC are abnormal in CD, but unlike DPD those patterns are not the result of VLPFC-induced hypoactivity.

Ex hypothesi the CD patient is extremely depressed. Evidence suggests that circuitry centred on the amygdala is affected, which means that online affective responses are flattened.

The role of the AIC is to monitor for changes driven by affective processing. It thus predicts for example that a typically positive event would be processed as positive. Thus, when that event is processed as negative or neutral, the AIC detects an error, signaled in the form of an anomalous experience. The patient is in the position being able to detect and signal changes in her affective responses, which take the form of unpredicted absences in bodily response. Thus *her lack of felt bodily response is processed as affectively significant* in the Cotard delusion with the result that she experiences it. Thus she does not feel neutral she feels miserable. Or as we might put it *she feels metamisery* because the role the AIC is to enable the person to feel the affective significance of bodily changes *including the absence of predicted changes*. In Cotard delusion the patient feels the affective significance the unpredicted absence of positive changes.

In DPD, by contrast, the patient does not feel the significance of bodily information because her AIC is inhibited and hypoactive.

Thus the difference between the two conditions may consist, not only in the fact that the Cotard delusion is a response to lower level prediction error, but the level in the predictive processing hierarchy at which predictions about bodily information are violated.

References

- Lin, Y.-T. (2015). Memory for prediction error minimization: From depersonalization to the delusion of non-existence—A Commentary on Philip Gerrans. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.

Visual Adaptation to a Remapped Spectrum

Lessons for Enactive Theories of Color Perception and Constancy, the Effect of Color on Aesthetic Judgments, and the Memory Color Effect

Rick Grush, Liberty Jaswal, Justin Knoepfler & Amanda Brovold

Many forms of visual adaptation have been studied, including spatial displacements (Heuer & Hegele 2008), spatial inversions and rotations (Heuer & Rapp 2011), removing or enhancing various colors in the visual spectrum (Belmore & Shevell 2011; Kohler 1963), and even luminance inversion (Anstis 1992). But there have been no studies that have assessed adaptation to an inverted spectrum, or more generally color rotation. We present the results of an adaptation protocol on two subjects who wore LCD goggles that were driven by a video camera, but such that the visual scene presented to subjects was color-rotated by 120°, so that blue objects appeared green, green objects appeared red, and red objects appeared blue (with non-primary colors being analogously remapped). One subject wore the apparatus intermittently for several hours per day for a week. The second subject wore the apparatus continually for six days, meaning that all his visual input for those six days was color rotated. Several experiments were run to assess the kinds and degrees of adaptation, including Stroop (1935), the memory color effect (Hansen et al. 2006), and aesthetic judgments of food and people. Several additional phenomena were assessed and noticed, especially with respect to color constancy and phenomenal adaptation. The results were that color constancy initially was not present when colors were rotated, but both subjects adapted so that color constancy returned. However, there was no evidence of phenomenal color adaptation. Tomatoes continued to look blue, subjects did not adapt so that they started to look red again. We found no reliable Stroop result. But there was an adaptation to the memory color effect. Also, interesting differences were revealed in the way color affects aesthetic judgments of food versus people, and differences in adaptation to those effects.

Keywords

Aesthetic judgments | Color constancy | Color phenomenology | Color rotation | Enactive perception | Inverted spectrum | Inverted spectrum thought experiment | Memory color effect | Phenomenal adaptation | Semantic adaptation | Stroop | Visual adaptation

Authors

[Rick Grush](#)
rick@mind.ucsd.edu
University of San Diego
San Diego, CA, U.S.A.

[Liberty Jaswal](#)

[Justin Knoepfler](#)

[Amanda Brovold](#)
abrovold@miracosta.edu
MiraCosta College
Oceanside, CA, U.S.A.

Commentator

[Aleksandra Mroczko-Wąsowicz](#)
mroczko-wasowicz@hotmail.com
國立陽明大學
National Yang Ming University
Taipei, Taiwan

Editors

[Thomas Metzinger](#)
metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Introduction

The idea of a subject whose visual experience is color inverted has been a philosophical mainstay at least since [Locke \(1975\)](#), and has fuelled a great deal of philosophical work on the nature of perception up to the present day. In psychology, testing the extent to which subjects' visual systems can adapt to alterations in visual input has likewise been a fruitful mainstay for over a century (for current research and references see [Heuer & Hegele 2008](#); [Heuer & Rapp 2011](#); [Belmore & Shevell 2011](#)). Despite these two facts, adaptation to an inverted spectrum has never been studied. Given that there is adaptation to a wide range of distortions to visual input, including spatial manipulations and spectral filtering, we speculated that it was conceivable that there might be adaptation, in some form or other, to color rotation. That is, if visual input was reworked such that tomatoes appeared blue, would subjects over time adapt so that tomatoes regained their normal (red) phenomenal appearance? And even if such a shocking result did not occur, might there nevertheless be adaptation to other color-relevant phenomena, such as color constancy, aesthetic judgments, or the memory-color effect?

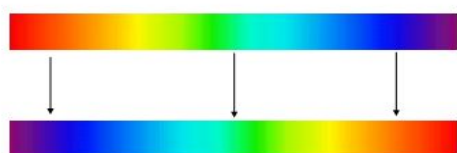


Figure 1: An inverted spectrum. The familiar red-to-violet spectrum laid out from left to right (top) compared to an inverted violet-to-red spectrum from left to right (bottom). Note that the central colors map to themselves or very similar colors.

We tested this as follows. First, rather than an inverted spectrum, we employed a *rotated spectrum*. An inverted spectrum is one in which one takes the usual red-to-violet spectrum and just reverses it to get a mapping from colors to colors. Red would map to violet, orange to blue, violet to red, and so on (see [figure 1](#)).

For two reasons we chose to systematically alter color input not with an inverted spectrum,

but with color rotation (see [figure 2](#)). In a 120° degree rotation, greens become reds, reds become blues, and blues become greens (see [figure 3](#)).

One reason to employ a rotation rather than an inversion is that in a rotation, all colors map to different colors, whereas in a spectral inversion, the middle of the spectrum maps to itself, and so not all colors differ. Also, with a 120° rotation, primary colors map to primary colors, and this was convenient for testing purposes. For instance, we wanted to test semantic adaptation via a Stroop task, and Stroop is difficult to test if one is dealing with non-canonical colors (since subjects pause to think of what seems to be the best name for the color: “... um, periwinkle?”). It was important that the text was colored in a primary color during baseline testing (before subjects' vision was color-altered), and that it continued to be presented in a primary color during testing while colors were altered.

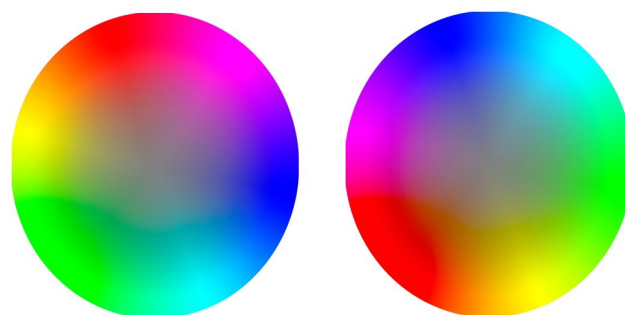


Figure 2: Color rotation. If colors on the left disk are mapped to colors that are 1/3 of a clockwise rotation (120°), then green (lower left) maps on to red (slightly left of the top); red maps to blue (middle right); and so forth. The disk on the right is a 120° color-rotated version of the disk on the left.

The rotation was implemented in the following way. A small video camera was mounted on a bike helmet, as were a pair of LCD goggles (see [figure 4](#)). The sides of the goggles were sealed against the wearer's face so that there was no peripheral visual input; subjects could only see what was displayed on the LCDs. The video camera output was run to a laptop computer, which ran software that color-rotated the

video feed and pushed the result to the LCD goggles. The subject would wear the bike helmet and carry a bag that held the laptop and battery packs for the laptop, camera, and goggles. In what follows we will refer to this as the *rotation gear*, or simply the *gear*. Moreover, we will use the expression *under rotation* to refer to the condition of having one's visual input color-rotated by the gear. In addition to rotating the spectrum (the intentional effect), the gear also eliminated binocular disparity since there was only one camera (which projected the same image to both eyes), and also impaired peripheral vision, since the camera's field of view was less than normal vision. As a result, while most normal activities were possible (if cumbersome), some, such as driving a motor vehicle, were not possible.

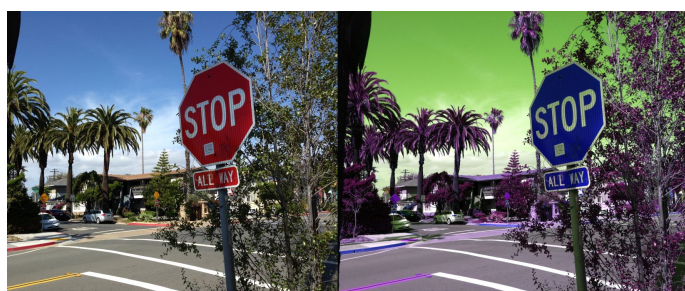


Figure 3: A typical outdoor scene color-rotated by 120°.
(<http://www.open-mind.net/videomaterials/grush-color-video>)

There were only two subjects, and both were investigators in this project (and authors of this paper). This was for practical reasons. First, we anticipated that UCSD Internal Review Board would be reluctant to grant human subjects approval, based both on the length and significant inconvenience of the protocol and also because this particular protocol had never been attempted before, and so, for example, there was no precedent concerning whether such a regimen might cause long-term damage to participants' color vision. And even if approval were obtained, we anticipated that finding volunteers for such a protocol would be difficult, and, even if we did find them, the small grant we were operating with did not give us the resources to appropriately compensate them. Employing two of the investigators as subjects solved these problems. Because investigators

were aware of the protocol and risks, and were intimately familiar with the relevant equipment and potential problems and so in a position to more accurately assess conditions under which the protocol should be aborted, some of the concerns were eased, and approval was eventually granted. Moreover, investigators were willing to put themselves through the arduous protocol.

Subject 1 (RG) wore the gear intermittently, in several multi-hour sessions per day for a week. Two reasons for an initial intermittent protocol were, first, that it would allow us to trouble-shoot the equipment—if it had to be shut down for a few hours for tinkering this would not interfere with the protocol, and we wanted to ensure that everything was functioning smoothly before subject JK's continuous protocol began. And, second, we wanted to compare the results of a subject who wore the gear intermittently with the results of one who wore it continually. Subject JK wore the gear continually for six straight days, meaning that he had no unrotated visual input for that entire period. He slept with a blindfold and showered with closed eyes in the dark, but otherwise wore the gear at all times.



Figure 4: The camera and goggles used.

Our high-level goals were to assess phenomenal and semantic adaptation. Phenomenal adaptation would manifest as a return to normalcy such that under rotation tomatoes would start to look red again, the sky would appear blue, and so forth (this is discussed in more detail later on). It might also manifest as a gaining of color constancy under rotation. These would be parallel to adaptation to spatial inversion in which, after adaptation objects begin to

look right-side-up again. We assessed this in several ways. One was the memory-color effect. A second was aesthetic judgments. A final method was subjective report: subject JK kept a hard-copy journal in which he wrote observations, and RG had a digital voice recorder that he used for similar purposes. There were also audio recordings made during the testing periods, as well as when JK finally removed the rotation-gear.



Figure 5: Images of people and food, in normal color, and with a 120° color rotation, for purposes of illustration only—none of these was in the stimuli set used in the experiment. The food items are French toast with maple syrup and strawberry confit (middle row), and chicken salad with guacamole (bottom).

Semantic adaptation would manifest as a remapping of color terms to their “correct” referents. So, for example, when first putting on the

gear, if a subject were asked to pick up the “blue block,” they would pick up a block that was in fact red. Would subjects semantically adapt such that the word “red” was immediately semantically connected to the red block, despite the fact that the block was presented as blue through the rotation gear? This was assessed via subjective report and Stroop. In the following sections we discuss each investigated phenomenon, the results we found, and their implications.

For all experiments there were four times at which trials were run: i) *pre-rotation*, in which trials were run after the subject initially put on the gear but the colors were not yet rotated; ii) *early rotation*, wearing the gear and first color-rotating visual input; iii) *late rotation*, at the end of the time during which the subject was wearing the gear and had time to adapt; and iv) *post-rotation*, while the subject was still wearing the gear, but colors were not rotated. The reason for running trials (i) and (iv) while wearing the gear without rotating the colors, rather than just through normal vision without the gear, was to control for effects possibly due only to the fact that visual input was going through a camera and LCD goggles. For some experiments there were additional times at which the trials were run, as will be explained below.

A final note on methodology. A number of factors distinguish the current study from an appropriately run and controlled psychological experiment. The small n and the fact that both subjects were also investigators in the study are perhaps the two most significant differences. These limitations were forced by a variety of factors, including the unusual degree of hardship faced by subjects, our relatively small budget, and the fact that this protocol had never been tried before. Because of these limitations, the experiments and results we report here are intended to be taken only as *preliminary* results—as something like a pilot study. Even so, the results, we believe, are quite interesting and suggestive.

2 Color, color constancy, and enactive vision

According to proponents of enactive perception, perceptual experience amounts to relevant beha-

vioral skills (O'Regan & Noë 2001; Noë 2004, [this collection](#)).

To be a perceiver is to understand, implicitly, the effects of movement on sensory stimulation [...]. An object looms larger in the visual field as we approach it, and its profile deforms as we move about it. A sound grows louder as we move nearer to its source. Movements of the hand over the surface of an object give rise to shifting sensations. As perceivers we are masters of the sort of sensory dependence [...]. We spontaneously crane our necks, peer, squint, reach for our glasses, or draw near to get a better look (or better handle, sniff, lick, or listen to what interests us). The central claim of what I call the enactive approach is that our ability to perceive not only depends on, but is constituted by, our possession of this sort of sensorimotor knowledge. (Noë 2004, pp. 1–2)

The enactive approach correctly predicts that there will be adaptation to certain kinds of spatial distortion to visual input (Noë 2004). The idea is that if perception is a matter of learning sensorimotor contingencies, then though these contingencies can be disrupted via altering the spatial features of the input, the new contingencies can be learned (they are just a different set of contingencies, after all) and when this happens, perceptual input will seem normal again.

Noë (2004) boldly claims that not just spatial features, but even color phenomenology might be explained on enactive principles.

Our ability to perceive [a] wall's color depends on our implicit understanding of the ways its apparent color varies as color-critical conditions vary. At ground, our grasp of these dependencies is a kind of sensorimotor knowledge. We can distinguish two different kinds of sensorimotor dependencies [...]. Crucially, the perceptual experience of color depends on the perceiver's knowledge of both kinds of sensorimotor patterns.

Movement-dependent sensorimotor contingencies are patterns of dependence between sensory stimulation, on the one hand, and movements of the body, on the other [...].

[O]bject-dependent [...] sensorimotor contingencies [...] are patterns of dependence between sensory stimulation and the object's movement, or the object's changing relation to its environment. (Noë 2004, pp. 129–130)

Accordingly, our protocol speaks as directly to the enactive account of color as inverting prisms speak to an enactive account of vision's spatiality. Notice that the technological apparatus of color rotation comes into play *after* both sorts of patterns have manifested (with the exception of eye movements, which happen after the rotation is effected, though this fact does not change any of the contingency patterns). What this means is this: suppose that there is a particular way that a red surface changes its reflectance properties both as we move around it (movement-dependent), and also as it changes relevantly with respect to the environment (object-dependent). Call this pattern of changes Pattern R. And suppose that the same is true for a blue surface, meaning that it has a different, but characteristic pattern of changes we can call Pattern B. To get a specific example, let's suppose that the red surface gets brighter when it gets angled upwards, but the blue surface does not (the details of these patterns don't matter for purposes of illustration, all that matters is that there are such patterns, and that they differ for different colors). Whatever the patterns R and B are, they occur whether anyone is wearing rotation gear or not. But after the rotation gear is involved the surface that is behaving according to Pattern R will be presented to the subject with stimulation from the blue part of the spectrum, and the surface that is behaving according to Pattern B will be presented with stimulation from the green part of the spectrum. For instance, the subject will see the *apparently blue* surface getting brighter as it is angled upwards, which

is Pattern R, because the apparently blue surface is actually red, and hence behaves according to Pattern R.

So the question is: after experience with red surfaces, which behave according to objective patterns appropriate to red surfaces (Pattern R), but are presented through the goggles with blue parts of the spectrum, will these surfaces start to look red again? This is clearly the prediction that is made by the enactive theory of color, since to appear red just is to behave according to Pattern R on this theory, and the surfaces that are being presented with light from the blue parts of the spectrum are behaving according to Pattern R. The enactive theory makes this prediction for both color constancy and color phenomenology. We will discuss color constancy first.

We did not test color constancy in any controlled way, but the subjective reports are quite unmistakable. Subject RG noticed that upon first wearing the rotation gear color constancy went “out the window.” To take one example, in normal conditions RG’s office during the day is brightly lit enough that turning on the fluorescent light makes no noticeable difference to the appearance of anything in the office. But when he turned the lights on after first donning the gear, everything had an immediate significant change of hue (though not brightness). He spent several minutes flipping the light on and off in amazement. Another example is that he also noticed that when holding a colored wooden block, the surfaces changed their apparent color quite noticeably as he moved it and rotated it, as if the surfaces were actively altering their color like a chameleon. This was also a source of prolonged amusement. However, after a few days the effect disappeared. Turning the office light on had little noticeable effect on the color of anything in his office, and the surfaces of objects resumed their usual boring constancy as illumination conditions or angles altered.

Subject JK reported the same thing: an initial period in which the apparent colors of objects shifted widely with changes in illumination conditions or viewing angles, followed after a day or two with the restoration of color con-

stancy such that those same changes had no effect on apparent color.

There was one difference, though, between RG and JK. While RG’s perceptual system gained the capacity for color constancy under rotation, he never lost color constancy in normal conditions. After the first few days of intermittently wearing the gear, objects had stable apparent colors whether he wore the gear or not. Though of course the colors that were stable were different in the two conditions. JK (who wore the gear continuously for six days) also gained color constancy under rotation, but lost it for normal conditions, as was apparent at the end of his trial when he removed the gear. Indeed, almost immediately after he removed the gear and was seeing things without rotation for the first time in six days, he spent several minutes flipping a light switch on and off and marvelling as the apparent color of everything in the room changed at his command (while no one else in the room noticed anything). So while RG’s visual system became, so to speak, bi-constant, because he switched back and forth between rotated and non-rotated visual input, JK’s visual system, because it was exclusively rotated for six days, gained color constancy under rotation, but temporarily lost normal color constancy. Normal constancy returned for JK within a few hours after he stopped wearing the gear.

These results are precisely what the enactive theory of color constancy would predict. Initially the kinds of patterns of color-relevant change exhibited by objects in the environment was different from what the visual system had come to expect, both in terms of changes in environmental conditions generally (e.g., switching on a fluorescent light), and movement specific changes (e.g., walking around them or rotating them in hand). The sensorimotor contingencies changed, and as a result color constancy was disrupted. But after a period of time during which these new dependencies were, presumably, learned, color constancy was restored.

A more convincing protocol would be one in which there were control subjects whose visual input was rotated, but who were not active in their color environment. Such a protocol

would be difficult to implement. Wearing rotating equipment for six days is quite difficult. If you were then to disconnect visual input from overt behavior on top of that, this would become extremely burdensome for test subjects. This could be done either by having the video camera not moving at all, or moving randomly; or by recording the video from a rotated subject as they are actively exploring their environment, and simply play that video back to control subjects, so that their visual input would not change at all as a consequence of their own actions. But even though there was no control of this sort in our protocol, it is safe to say that the proponent of an enactive theory of color constancy should be encouraged by this result.

But what about color adaptation? Did red surfaces start to look red again? The results here are less encouraging for the enactive theorist. With one interesting and suggestive observation to be discussed shortly, we found nothing that suggested color adaptation. As assessed by subjective report, stop signs continued to appear blue, the sky green, and broccoli red throughout for both subjects.

Though this is not the result that the enactive theorist would hope for, it isn't entirely conclusive. First, it may have been the case that a protocol of longer than six days would have resulted in phenomenal adaptation. Six days may simply not have been long enough to learn the new relevant sensorimotor contingencies. This is certainly possible. But it should be noted that all other sorts of visual adaptation (to spatial inversions, spectral filtering etc.), including our own result with adaptation to color constancy, occurred in less than six days.

Second, the gear does in fact introduce a lot of artefacts besides just the change in presented colors. Artefacts are introduced by the digitization of the image and its presentation through LCD goggles. So proponents of the enactive theory of color needn't jump off a building just yet. They may maintain that because of these artefacts, the process of relearning the needed sensorimotor contingencies was somehow short-circuited.

But there are a couple of considerations that suggest that the result is not so easily dis-

missed. First, whatever artefacts the gear did introduce were not such as to make any difference to normal color vision. If one wears the gear without color-rotation, things appear to be in their normal colors. This seems to suggest that whatever patterns of change account for our ability to see normal things in their normal colors, they are not significantly compromised by whatever artefacts the gear introduced. This would seem to remove some of the motivation from the suggestion that adaptation did not occur because of artefacts introduced by the gear. Second, the gear clearly maintained enough information about patterns of color change that adaptation to color constancy occurred fairly quickly. Again, this suggests that much or all of whatever is important in patterns of change relevant to color perception is preserved by the gear. If it was not, adaptation to color constancy should not have occurred. The one thing not preserved by the rotation gear was the sensorimotor-*independent* feature of which retinal cells were stimulated when various surfaces were in view. And that one feature seems to be the best candidate for the determinant of apparent color, given that everything else changed but apparent color did not.

We mentioned above that there was an interesting pair of events that, while not quite amounting to phenomenal adaptation, are at least very suggestive. On two occasions late into his six-day period of wearing the gear, JK went into a sudden panic because he thought that the rotation equipment was malfunctioning and no longer rotating his visual input. Both times, as he reports it, he suddenly had the impression that everything was looking normal. This caused panic because if there was a glitch causing the equipment to no longer rotate his visual input, then the experimental protocol would be compromised, and the value of his days of sacrifice in the name of science and philosophy would have been significantly diminished.

However, the equipment was not malfunctioning on either occasion, a fact of which JK quickly convinced himself both times by explicitly reflecting on the colors that objects, specifically his hands, appeared to have: "OK, my hand looks purplish, and purple is what it

should like under rotation, so the equipment is still working correctly.”

Prima facie there seems to be a clear difference between i) a tomato looking “normal” because it now appears phenomenally to be red; and ii) a tomato looking “normal” because though it appears blue, one is now used to tomatoes appearing blue, that is, blue no longer appears unusual. JK’s situation was a case of (ii), but the lack of a sense of novelty of strangeness made him briefly fear that he was in a (i) situation. He described it as a cessation of a “this is weird” signal.

Even though (i) and (ii) seem to be quite different, the phenomenon is suggestive. It indicates that there is definitely a stage in which the subject requires explicit reflection to discern (i) from (ii). This might lead one to speculate that this stage might signal an early stage of genuine color adaptation (we will discuss this further in the final section).

But it could also be an initial stage of a very different possibility, one discussed by Noë himself:

[...] the strongest [inverted spectrum] arguments ask us to consider the possibility in the first person. At stage 1 I am inverted. At stage 2, I get used to the inversion. I realize things now look color-inverted compared to the way they used to look, and I use this knowledge to guide my correct use of words. I get really good at acting normal. At stage 3, I suffer amnesia and forget that things ever looked different. The point of this thought experiment is that it suggests a reason to believe that things are now different with me with respect to my color experience, even though I am now unable to report those differences. (2004, p. 94)

While JK never suffered from amnesia, his two episodes suggest that it is possible to at least go some distance down precisely this path.

3 The memory-color effect

If subjects are given control over the hue of an image and asked to adjust it until it appears

grey scale there is an interesting effect. If the image is of an object with a salient prototypical color (such as bananas, which are saliently and prototypically yellow), subjects will judge that it is grey scale when in fact the hue is slightly in the direction opposite to that of the standard color. So, for example, an image of a banana will be judged to be grey scale when it is in fact just slightly periwinkle. This is the memory-color effect (Hansen et al. 2006). One possible explanation of this effect is that when the image actually is grey scale, subjects’ top-down expectations about the usual color make it appear (in some way or another) to be slightly tinted in that hue. So when the image of the banana is actually completely grey scale subjects judge it to be slightly yellow. The actual color of the image must be slightly in the direction opposite yellow (periwinkle) in order to cancel this top-down effect and make the image appear grey. This is the memory-color effect.

We expected that after a period of wearing the gear this effect would diminish or even reverse. The reasoning was that if the rotated experience either just disrupted the association of the objects with their prototypical color, or even established a different prototypical color, the effect would be compromised.

Stimuli used in our trials were images of a banana, tomatoes, broccoli, a fire engine, a school bus, a stop sign, and a Starbucks logo. We wanted examples of natural objects as well as artefacts with strong color associations. We also used squares and circles, which have no obvious prototypical color, as controls. The hue of the initial image presentation was random, and subjects were to adjust the hue until the image appeared completely grey scale.

There were technical issues with RG that prevented his data from being usable. But subject JK’s results were in fact what we expected. Pre-rotation, JK’s results were normal. He judged the stimuli to be grey when in fact they were, on average, 3.5% saturated in the direction of the opponent color. JK was quite consistent with this except for one stimulus condition, broccoli, which he actually exhibited the opposite of the expected pattern. He judged it to be grey when it was about 1% saturated in

its normal color. The effect was robust, and we have no idea why. We suspect that it was some sort of artefact connected to that stimulus being presented through the rotation gear, but we don't know for sure.

Post-adaptation JK's assessments were, on average, about 0.5% saturation, again in the direction of the opponent color. Meaning that he still exhibited the effect, but its magnitude was lessened. This could mean one of two things: i) the rotation protocol disrupted the usual associations of colors and objects, and so all stimuli ended up being treated by his perceptual system just as controls, that is, as objects with no salient associated color; and ii) JK partially adapted to objects' new prototypical colors. Unfortunately the result we got, in which assessments were on average very close to grey, is consistent with both. The alteration was consistent with both a movement towards grey and a movement towards the canonical color, since both are in the same direction from a spot opposite the canonical color. But one consideration that speaks in favor of (ii) is that the broccoli stimuli, post-adaptation, did not move towards grey, but in fact were judged to be grey when they were even more green than in the pre-rotation trial. That is, the broccoli stimuli judgments moved not in the direction of grey, but in the direction of canonical color, and by about the same amount as the other stimuli moved in that direction: 2.5%.

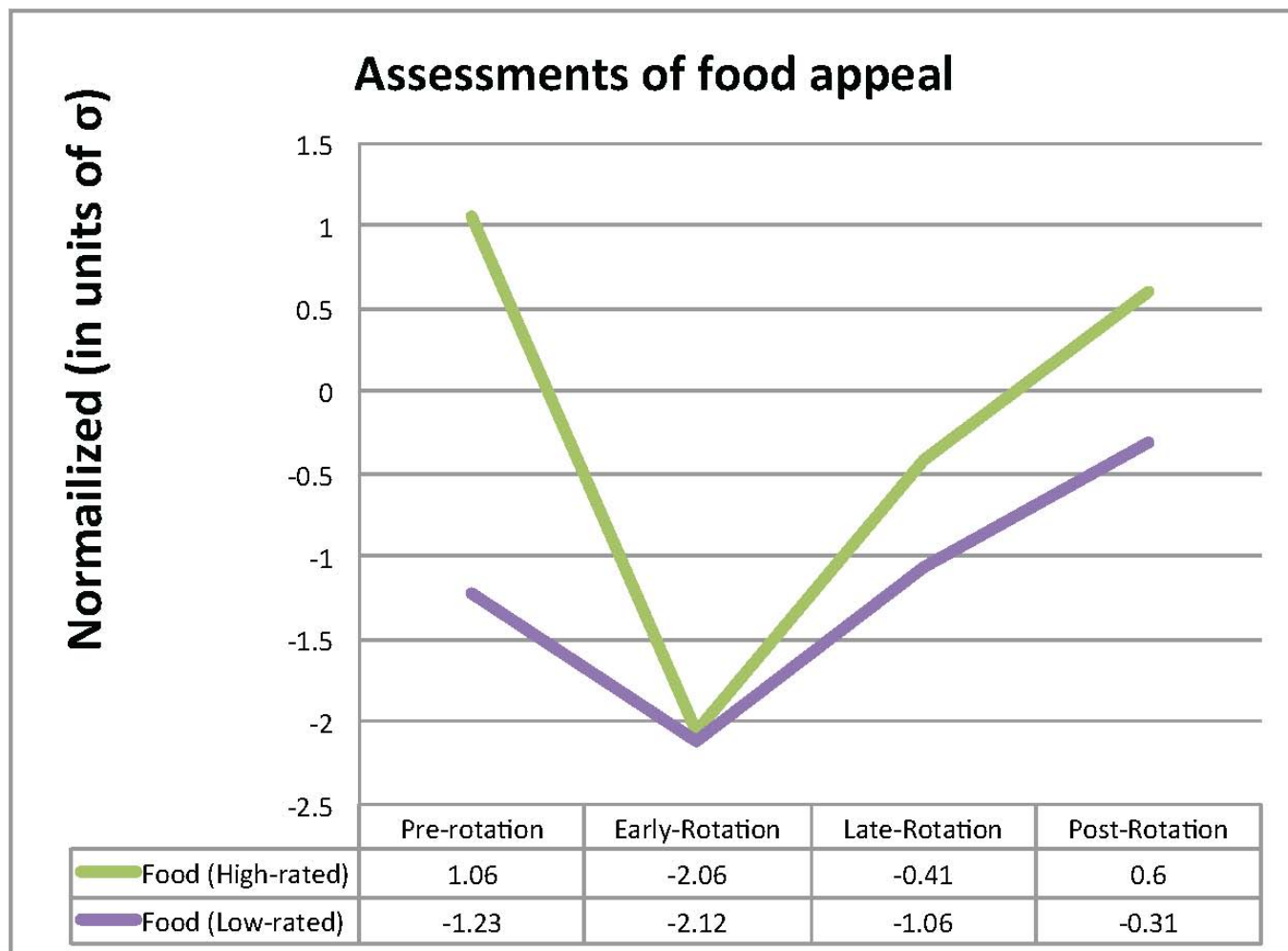
Our result in the memory-color effect and the two occasions in which JK panicked are consistent. Both effects would seem to result from a re-aligning of the salient prototypical color of objects that have a salient prototypical color. The adaptation of the memory-color effect suggests that the experience of being color-rotated lessened the extent to which top-down effects associated certain objects with their actual prototypical colors, and perhaps even started associating them with new, different prototypical colors. And the lack of the "this is weird" signal when viewing his purple hand also suggests that the old prototypical look of his hand was being supplanted with a new prototypical look. And it also suggests not just that the old prototypical color association was being

lessened, but that a new one was emerging. What failed to look weird was his *purple* hand—not just a hand in any non-flesh color, but in *purple*. We did not test this, but it seems quite unlikely that had his hand suddenly looked bright red he would have similarly experienced a loss of the "this is weird" signal. This is speculative, but it at least suggests, consistently with the broccoli effect discussed above, that the result we saw with the memory-color effect was not just a loosening of the old associations, but the emergence of new ones.

Moreover, the adaptation of the memory-color effect appears to have been *general*. The items we used for testing in the memory color effect fell into two groups: first, there were those items, such as Starbucks logos and bananas, which were such that items of that type were observed at least sometimes during the period of rotation; and second there were others, such as fire trucks and baby chicks, which were not observed by the subject under rotation. If what was being altered by rotated experience was just the specific associations of colors with experienced objects, then we should have found different results for these two groups of objects. Bananas and Starbucks logos would be subject to adaptation with respect to the memory-color effect, and fire trucks and baby chicks would not. But this is not what we found. We found the memory-color effect was impacted for *all* tested objects, even those that had not been seen during rotation.

This suggests that the effect was general, meaning that the adaptation was manifested not as an alteration in some part of the perceptual system concerning its expectations about what bananas or other specific objects look like. Rather, the alteration appears to have concerned expectation about what *yellow things* generally look like. To put it in dangerously loose and anthropomorphic terms, some part of the system started cranking up the periwinkle knob when objects known to be yellow, like baby chicks, were seen. If this is indeed a *memory* effect (psychologist do, for some reason, call it the *memory-color* effect), then it suggests that some part of the system knew that baby chicks were supposed to be yellow, but was beginning to misremember what yellow was like.

Table 1: Assessments of food appeal. The graphs show the average scores for both subjects. For each subject, before testing each subject scored all images on a scale of 1–10, 10 being the most appealing. These scores were used to establish a normalization for each subject so that all scores could be re-expressed in terms of σ . They were also used to divide the dishes into high-rated and low-rated groups for each subject. Each subject then re-assessed each dish at four times during the experiment: pre-rotation, early-rotation, late-rotation, and post-rotation (see text for explanation). As can be seen, upon color rotation all dishes plummeted in their perceived appeal to a level below the score the unappealing dishes had before rotation. But after adaptation (late rotation) the low-rated dishes regained all of their appeal, and the high-rated ones regained a moderate amount.



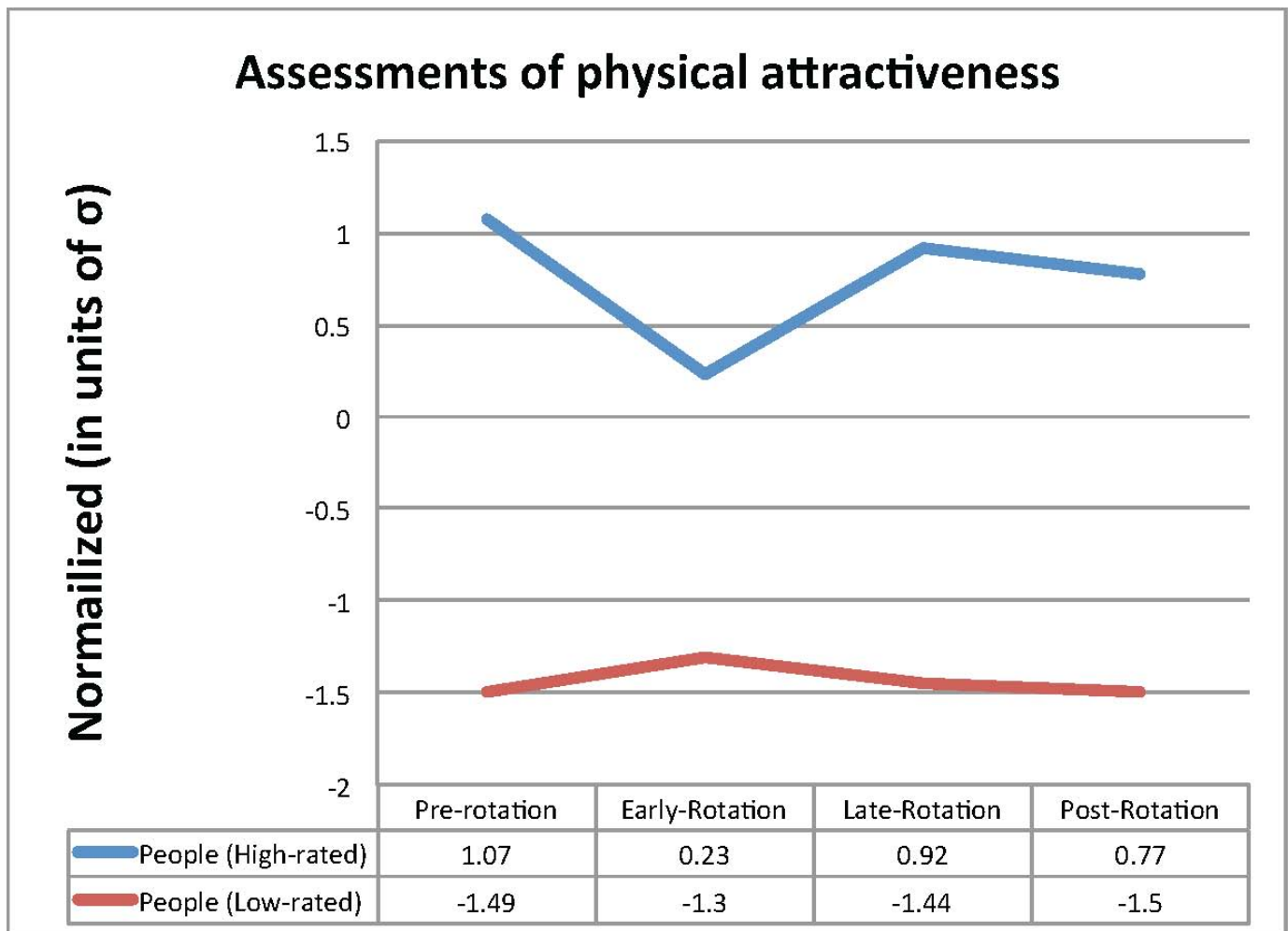
Recall where we left off in the last section. There was a suggestion to the effect that for a subject who underwent an inverted spectrum procedure, a cessation of the “this is weird” signal, together with a loss of memory of how things used to look might result in that subject’s inability to report differences between their current inverted experience and their prior non-inverted experience. Such a subject would of course verbally report that their phenomenology had adapted (or, if they were more reflective, they might also admit the possibility that their phenomenology was still inverted, but that

their memory was failing to make this fact apparent to them). To indulge in some wild speculation, the general nature of the adaptation to the memory-color experiment suggests that something along these lines might possibly happen if there was a longer adaptation period. We will return to this in the final section.

4 Aesthetic judgments

Color plays a large role in a variety of aesthetic judgments. Red broccoli doesn’t look terribly appealing as a food item, and Hollywood main-

Table 2: Assessments of physical attractiveness of people. The graphs show the average scores for both subjects. For each subject, before testing each subject scored all images on a scale of 1–10, 10 being the most attractive. These scores were used to establish a normalization for each subject so that all scores could be re-expressed in terms of σ . They were also used to divide the images into high-rated and low-rated groups for each subject. Each subject then re-assessed each image at four times during the experiment: pre-rotation, early-rotation, late-rotation, and post-rotation (see text for explanation). As can be seen, color rotation had much less effect on ratings of physical attractiveness of people than it did on the appeal of food. The low-rated group was essentially unchanged throughout the protocol. The high-rated group experienced a relatively small drop during early-rotation, but regained nearly all of it by late rotation.



tains that gentlemen prefer blondes. But to what extent are these judgments malleable with experience? Might the red broccoli start to look more appetizing if one has enough experience eating it? To assess these questions we ran two aesthetic judgment experiments. In one, we measured how appealing different dishes looked, and in another we measured judgments of physical attractiveness of people (figure 5). For the first we used images from a large cook-book with a wide variety of dishes. For the second, since both subjects were heterosexual males, we used pictures of adult women taken from the in-

ternet. In both cases we separated the stimuli into high-rated and low-rated groups so that we could assess whether color rotation had a differential effect. And in both cases we ran four trials: an unrotated baseline trial of judgments before beginning the rotation period; an early-rotation trial immediately upon wearing the rotation gear; a late-rotation trial at the end of the adaptation period; and finally a post-rotation trial normal color vision was restored. Subjects scored all stimuli with a 1–10 numerical rating. We used the normal vision early-rotation ratings of each subject both to sort stimuli into

subject-specific high-rated and low-rated groups, and to establish a normalization (average and standard deviation) so that we could translate all ratings provided by each subject in terms of σ , meaning that 0 was average, 1 was one standard deviation above average, -1 was one standard deviation below average, and so forth.

The results can be summarized quickly (see table 1). For dishes in the appealing group, their appeal ratings immediately dropped significantly from pre-rotation to early-rotation. The drop was, on average, 3σ , from an average of 1.06σ to an average of -2.06σ ! Some notes from JK's diary speak to the effect of color rotation on food appeal:

Spinach looks a glossy, poisonous red (it's the texture and the color that look really nasty together, I think).

I genuinely lost my appetite at the sight of four enormous bowls of glossy red salad, pale pink cheese, blue kidney beans, deep blue beets, bright red peas, etc. [...]

I've noticed that I don't anticipate food at all the way I normally do. Normally during the day I think about food, think about making food, think about eating food, and when I get a full plate of tasty looking food in front of me, I'm a very happy person. Now there's a real disconnect between the way food tastes and the way it looks, and I don't honestly find myself craving anything during the day. I get hungry, yes, but a full plate of food in front of me looks intensely neutral on the desirability scale.

At the end of the adaptation period, those high-rated dishes had regained half of their appeal, being judged on average about -0.41σ . Both RG and JK followed this pattern. Interestingly though, in the post-rotation rating, RG's ratings returned to within $.35\sigma$ of baseline for the appealing dishes (to $.71\sigma$ compared to an initial average rating of 1.06σ), while JK's ratings were lower, returning only to within 0.61σ of

baseline, from 1.09σ to 0.48σ . But this is still a very significant absolute gain. On average, both ratings returned to 0.60σ post-rotation. Again, from JK's diary after his return to normal vision and going to a salad bar:

The various greens were really intensely "green," the carrots looked like they were going to leap out of the buffet line. The onions (red onions) were the only vegetable that didn't surprise me with its vividness. I was horribly hungry; eating was immensely pleasurable. You have no idea how nice it is for food to look and taste right.

For the less-appealing group of dishes the ratings also dropped immediately from pre-rotation to early-rotation, on average by $.89\sigma$, from -1.23σ to -2.12σ . This group of dishes recovered to -1.06σ in late-rotation, just slightly above their pre-rotation ratings. In post-rotation this group of dishes rose to above their baseline ratings, to -0.31σ , nearly a full standard deviation above their pre-rotation baseline!

Interestingly, the overall pattern was that while the high-rated dishes started at 2σ , and low-rated dishes started at -1σ , they all dropped to the same -2σ immediately upon early-rotation. The color rotation just made everything, dishes that normally looked appealing as well as those that did not, look very unappealing. After adaptation at late-rotation, the lower-rated dishes recovered all of their perceived appeal (low though it was), and the higher-rated dishes recovered half. But it is worth remarking that the appealing dishes had a larger absolute gain after adaptation, suggesting that experience with appealing and unappealing food did have an effect on how each subject judged the appeal of foods based on color. Unappealing dishes returned to their normal level of unappeal, and the appealing dishes made significant gains towards looking as appealing as they had before.

As for the physical attractiveness assessments, the effect of color rotation was much less pronounced than it was in the case of food. We dubbed this the Star Trek effect (in honor of

Captain Kirk’s romantic liaisons with green alien women): attractive people are still pretty attractive whether their skin is blue, or green, or any other color.

The results are summarized in table 2. As in the case of food, we did an initial norming trial, and then separated stimuli into high-scoring and low-scoring groups. And we tested at the same four times: pre-rotation, early-rotation, late-rotation, and post-rotation. For the low-scoring stimuli, there was almost no difference across the four trials. The average rating in this group started at -1.49σ pre-rotation, and then actually (slightly) improved during early-rotation to -1.30σ . Late-rotation ratings fell to -1.44 , and then post-rotation ratings were at -1.50 . Basically, color rotation had no pronounced effect on this group, either before or after the adaptation period.

For the high-scoring stimuli the pattern was more interesting. Pre-rotation average was 1.07σ , which dropped to 0.23σ immediately at early-rotation. While this is indeed a drop, it is a *much* smaller drop than the corresponding condition with food, in which high-rated dishes dropped 3.00σ . High-rated people ratings dropped only about $\frac{1}{4}$ as much as high-rated food ratings upon rotation. Late-rotation the high-scoring people had nearly completely rebounded to 0.92σ , only 0.15σ less than their pre-rotation scores. Post-rotation the ratings dropped slightly to 0.77σ , meaning that the people were actually judged to be less attractive in their normal color, than they were when color-rotated, though the degree of drop was small.

The difference we found between food and people is consistent with studies involving less extreme color manipulation. It is well known that color has a very large effect on the appeal of food (Delwiche 2004; Zampini et al. 2007; Shankar et al. 2010). Existing studies of food-color effects typically involve less extreme color manipulations than we studied here, but the results are similar. With people, though skin color does have an effect on perceived attractiveness, structural features (e.g., facial bone structure, symmetry, body proportions) seem to be more significant (Barber 1995; Dixon et al. 2007).

5 Semantic adaptation

We were keen to investigate the extent to which there would be semantic adaptation. When a normal fluent English speaker hears “red” as part of a sentence, a host of cognitive and behavioral states and processes are invoked that are keyed to a certain phenomenal appearance. If I ask you to “hand me the red block” you can immediately and without overt reflection grasp the correct block, even if it is surrounded by other, differently colored blocks. To what extent would color-rotated subjects show semantic adaptation? Would a point be reached at which the sentence “hand me the red block” just as fluently resulted in the grasp of the block that was in fact red, but presented as blue?

To help facilitate semantic adaptation both subjects spent a good deal of time each day performing tasks requiring engagement with color vocabulary. The most relevant of which was the building game, in which subjects were given written or verbal descriptions for building constructions from colored blocks. The constructions required blocks of specific colors to be placed in specific locations and orientations. Instructions were given in terms of the blocks’ actual color. Success required that subjects select the correct blocks even though the color terms used mismatched their visual input.

Our main method of testing adaptation in a detailed way was Stroop (1935). For technical reasons we have data only for JK (RG’s trail exposed a problem with the interaction between the goggles and the computer display presenting the stimuli, which rendered those data unusable but did allow us to fix the issue so that we could collect valid data from JK). In one standard Stroop set-up, subjects are shown color words presented in colored text, the text either spelling out a color name, or being a neutral series of asterisks, such as ****. For example, subjects might see the word RED in blue text. The task is to name, as quickly as possible, the color of the text while ignoring the color named by the word. When presented with the word RED in blue text, the subject is to say “blue” as quickly as possible.

The standard result is that there is significant interference. Subjects are fastest and most accurate when the color of the text and the color word match, as in the word BLUE in blue text. When there is a mismatch, they are slower and commit more errors, especially errors in which the subject replies with the color word named by the text, and not the color of the text. Our hypothesis was this: with colors rotated by 120° but before a period of adaptation, subjects would show the same pattern as normal subjects in that they would have interference when the color named by the text mismatched the rotated color of the text. So for example the word RED in red text (which would be presented as blue through the goggles) would result in interference, but the word RED in green text (which would be presented as red) would not. But after wearing the gear for a period of time there would be some degree of semantic adaptation. This would manifest in two ways. First, there should be facilitation, or at least diminished interference, when the color named is the same as the actual color of the text, even though it is presented in a different color. And second, there should now be interference, or less facilitation, when the color named is the same as the color in which the word is presented, because that color would be different to the color the word actually is.

The results were that we found no significant effect in either direction. This was disappointing, but it is consistent with the subjective reports of both RG and JK. They both remarked that it quickly became easy to do the appropriate translation and, for instance, grab the green block when one was instructed to “grab the green block” despite its being presented as red through the goggles. But even near the end of their adaptation periods, it still felt like it was an active (though fast and easy) translation, and not a pre-reflective semantic connection between “green” and the actually green objects.

6 Discussion

All of the experimental results reported here should be treated as pilot study results. Our n

was very small, either 1 or 2 depending on the experiment, and investigators were themselves experimental subjects. Both of these are significant limitations. Nevertheless we’re confident that any follow-up studies, with larger n , with more subjects who are not investigators, and perhaps with better equipment than we had available (for example, a camera with higher resolution), will yield results consistent with ours. We hope that any group that chooses to follow up will have an easier time than we did. To that end, we can report that post experiment both subjects’ vision returned to normal, including color discrimination (which we assessed a few days post gear), and so perhaps with verification in hand that the protocol is safe, it will be a little easier for others to get approval for use of human subjects.

The experiment as we have described it was designed to assess, among other things, phenomenal adaptation. We’ve acted so far as though what this means is obvious. But we’re now in a position to see that it isn’t obvious at all. For anyone who believes in qualia (and [Dennett 1988](#) is right when he says that most people do, philosophers or scientists, whether they realize it or admit it or not), then the idea would be straightforward. Phenomenal adaptation would occur when, for instance, tomatoes start causing red qualia again, even if the subject is wearing the rotation gear. There are alternatives to the qualia theory. The enactive approach offers one: the idea would be that phenomenal adaptation is nothing but *enactive adaptation*, that is, learning new sets of sensorimotor (and related) contingencies. Our results, especially the lack of any straight-forward phenomenal adaptation, though far from decisive put at least a little pressure on the enactive view, however. Another approach (e.g., [Dennett 1988](#)) would be to first cash phenomenology out in terms of inner discriminatory states that are tied to various reactive potentials. Described in this broad way, the enactive approach would be a very special case. The reactive potentials would include behavioural dispositions and possibilities, and even predictive possibilities, but also aesthetic reactions, emotional reactions, cognitive reactions, and so forth. As [Dennett](#)

puts it, the mistake made by the believer in qualia is the mistaken belief that

[...] we can isolate the qualia from everything else that is going on—at least in principle or for the sake of argument. What counts as the way the juice tastes to *x* can be distinguished, one supposes, from what is a mere accompaniment, contributory cause, or byproduct of this “central” way. One dimly imagines taking such cases and stripping them down gradually to the essentials, leaving their common residuum, the way things look, sound, feel, taste, smell to various individuals at various times, independently of how those individuals are stimulated or non-perceptually affected, and independently of how they are subsequently disposed to behave or believe. (1988)

On this view, JK (and to a lesser extent RG) may have been part way down the path to the only thing that would legitimately count as phenomenal adaptation, namely, changes in the way that some inner discriminatory ability is evoked and what its various consequences are. As we found, aesthetic judgments had started to adapt, and even the memory-color effect had begun to adapt in a way that speculatively may have been a reflection of alterations in what canonical colors were supposed to look like. Moreover, JK was losing his “this is weird” signals. And though we did not find evidence of semantic adaptation, it would be quite surprising, given humans’ ability to learn new languages and dialects, if after a more extended period of time semantic adaptation did not occur.

Whatever the details, for purposes of this experiment, we don’t feel compelled to take a detailed stand on any of this. We would have been satisfied with subjective report of phenomenal adaptation, and then left it to further philosophical and even psychological investigation to unpack what this could mean. Nevertheless, the adaptation to color constancy and the memory-color effect, as well as the loss of the “this is weird” signal, are all suggestive results

that we hope will help move debate in the relevant fields forward.

Acknowledgements

We would like to thank Paul Churchland for assistance in constructing the physical equipment. His job mounting the video camera and LCD goggles on the bike helmet was fantastic. We would also like to thank Eileen Cardillo and Tanya Kraljic for assistance with the Stroop experiments. We also received excellent feedback and advice from many people, including Stuart Anstis, Vilayanur Ramachandran, Pat Churchland, Paul Churchland, two referees, and many others. This work was supported by a grant from the UCSD Academic Senate. It is dedicated to the memory of Liberty Jaswal, one of the investigators who was originally intended to be the second subject, who died tragically just before we were about to begin data collection.

References

- Anstis, S. (1992). Visual adaptation to a negative, brightness-reversed world: Some preliminary observations. In G. Carpenter & S. Grossberg (Eds.) *Neural networks for vision and image processing*. Cambridge, MA: MIT Press.
- Barber, N. (1995). The evolutionary psychology of physical attractiveness: Sexual selection and human morphology. *Ethology and Sociobiology*, 16 (5), 395-424. [10.1016/0162-3095\(95\)00068-2](https://doi.org/10.1016/0162-3095(95)00068-2)
- Belmore, S. C. & Shevell, S. K. (2011). Very-long-term and short-term chromatic adaptation: Are their influences cumulative? *Vision Research*, 51 (3), 362-366. [10.1016/j.visres.2010.11.011](https://doi.org/10.1016/j.visres.2010.11.011)
- Delwiche, J. (2004). The impact of perceptual interactions on perceived flavor. *Food Quality and Preference*, 15 (2), 137-146. [10.1016/S0950-3293\(03\)00041-7](https://doi.org/10.1016/S0950-3293(03)00041-7)
- Dennett, D. (1988). Quining qualia. In A. Marcel & E. Bisiach (Eds.) *Consciousness in modern science*. Oxford, UK: Oxford University Press.
- Dixon, B. J., Dixon, A. F., Morgan, B. & Anderson, M. (2007). Human physique and sexual attractiveness: Sexual preferences of men and women in Bakossiland, Cameroon. *Archives of Sexual Behaviour*, 36 (3), 369-375. [10.1007/s10508-006-9093-8](https://doi.org/10.1007/s10508-006-9093-8)
- Hansen, T., Olkkonen, M., Walter, S. & Gegenfurtner, K. R. (2006). Memory modulates color appearance. *Nature Neuroscience*, 9 (11), 1367-1368. [10.1038/nm1794](https://doi.org/10.1038/nm1794)
- Heuer, H. & Hegele, M. (2008). Constraints on visuo-motor adaptation depend on the type of visual feedback during practice. *Experimental Brain Research*, 185 (1), 101-110. [10.1007/s00221-007-1135-5](https://doi.org/10.1007/s00221-007-1135-5)
- Heuer, H. & Rapp, K. (2011). Active error corrections enhance adaptation to a visuo-motor rotation. *Experimental Brain Research*, 211 (1), 97-108. [10.1007/s00221-011-2656-5](https://doi.org/10.1007/s00221-011-2656-5)
- Kohler, I. (1963). The formation and transformation of the perceptual world. *Psychological Issues*, 3 (4), 1-173.
- Locke, J. (1975). *An essay concerning human understanding*. Oxford, UK: Clarendon Press.
- Noë, A. (2004). *Action in perception*. Cambridge, MA: MIT Press.
- (2015). Concept pluralism, direct perception, and the fragility of presence. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- O'Regan, J. K. & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24 (5), 939-973. [10.1017/S0140525X01000115](https://doi.org/10.1017/S0140525X01000115)
- Shankar, M. U., Levitan, C. A. & Spence, C. (2010). Grape expectations: The role of cognitive influences in color-flavor interactions. *Consciousness and Cognition*, 19 (1), 380-390. [10.1016/j.concog.2009.08.008](https://doi.org/10.1016/j.concog.2009.08.008)
- Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18 (6), 643-662. [10.1037/h0054651](https://doi.org/10.1037/h0054651)
- Zampini, M., Sanabria, D., Phillips, N. & Spence, C. (2007). The multisensory perception of flavor: Assessing the influence of color cues on flavor discrimination responses. *Food Quality and Preference*, 18 (7), 975-984. [10.1016/j.foodqual.2007.04.001](https://doi.org/10.1016/j.foodqual.2007.04.001)

What Can Sensorimotor Enactivism Learn from Studies on Phenomenal Adaptation in Atypical Perceptual Conditions?

A Commentary on Rick Grush and Colleagues

Aleksandra Mroczko-Wąsowicz

Grush et al. present a pilot study on visual adaptation to a remapped color spectrum. Their preliminary results, being far from conclusive, only partially support the hypothesis that there might exist a form of adaptation to color rotation and color constancy. Proving such flexibility in color vision would substantiate the investigators' attempt to localize their research outcomes in the context of philosophical theories of enactive perception. In spite of some limitations, the study exhibits a worthy and novel approach to the old question of color inverted experience, intended to provide an interdisciplinary account that is both empirically sensitive and philosophically potent. For the progress of the current investigation it would be constructive not only to conduct empirical follow-up studies, but also to conceptually refine the notion of "phenomenal adaptation", which is the central phenomenon studied here.

Based upon a distinction between phenomenal conservatism that accepts only perceptual phenomenology with sensory contents and phenomenal liberalism that acknowledges higher-level contents of perception and cognitive phenomenology, I differentiate between adaptation of the sensory sort and adaptation in the cognitive aspects of experience.

This distinction is used to highlight two different ways of understanding the notion of "phenomenal adaptation", exhibited by the target article and this commentary. Grush et al. seem to suggest that phenomenal and (non-phenomenal) semantic adaptation are different forms of a more general phenomenon of adaptation. However, they do not give any explicit example of the genus of adaptation of which these types are a species. I contend, in turn, that there is no need to produce such subclasses of the notion; semantic adaptation involving higher-level non-sensory states may also be understood as phenomenal. This follows from phenomenal liberalism. I argue that what is being processed in the course of phenomenal adaptation is phenomenal character understood in an expansive way that includes high-level contents. The claim may have an important effect on related empirical work. As a result, enactive sensorimotor adaptation does not have to be seen as adaptation of the sensory sort, but as adaptation in the cognitive aspects of experience, such as altered expectations, or beliefs about or sensitivity to kinds of objects encountered in perceptual experience. This phenomenally liberal reading would provide an appropriately more capacious notion than the adaptation of the sort offered by Grush et al.

Finally, I claim that the lessons for enactive theories of color perception may be expanded beyond the implications of the color rotation study. This is demonstrated by turning to confirmatory and challenging cases of atypical perceptual conditions and color modifications, such as synesthetic color experiences.

Keywords

Adaptation | Color inversion | Color vision | Constancy | Enactivism | Inverted spectrum | Perceptual experience | Phenomenal character | Sensorimotor | Sensorimotor contingency | Synesthesia

Commentator

Aleksandra Mroczko-Wąsowicz
mroczko-wasowicz@hotmail.com
國立陽明大學
National Yang Ming University
Taipei, Taiwan

Target Authors

Rick Grush
rick@mind.ucsd.edu
University of San Diego
San Diego, CA, U.S.A.

Liberty Jaswal

Justin Knoepfler

Amanda Brovold
abrovold@miracosta.edu
MiraCosta College
Oceanside, CA, U.S.A.

Editors

Thomas Metzinger
metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Introduction

Philosophical thought experiments focusing on different kinds of visual spectrum manipulation and color inversion were initiated with John Locke's hypothetical case of strawberries producing visual experiences of cucumbers (Locke 1689/1979). They still influence not only philosophical theories of color perception and color qualia inversion (e.g., Shoemaker 1982; Clark 1985; Levine 1988; Block 1990; Casati 1990; Broackes 1992; Hardin 1993; Tye 1993, 2000; Nida-Rümelin 1993, 1996; Byrne & Hilbert 1997; Hurley 1998; Hilbert & Kalderon 2000; Cohen 2001; Myin 2001; McLaughlin 2003; Noë 2005; Churchland 2005; Macpherson 2005; Cohen & Matthen 2010; Burge 2010; O'Regan 2011), but also psychological research on the various ways in which our conscious experience can be modified and adapted to changes in visual input, such as space or luminance inversion (Heuer & Rapp 2011; Anstis 1992), or removing or enhancing colors (Belmore & Shevell 2011). However, a systematic interdisciplinary study on adaptation to an inverted or rotated color spectrum has been lacking until now.

The target article aims to lay the foundation for this, by presenting an experimental pilot study, along with some preliminary results and a brief discussion of its theoretical implications. Like many other pilot studies, it faces some limitations. These are: the small number of subjects tested; experimenters acting as test persons; and a complete lack of control conditions in the experimental protocol. The investigators are aware of these constraints and provide convincing reasons for the choices and strategy, e.g., their use of a novel, unpredictable, long-lasting, and inconvenient test protocol. Despite some difficulties relating to both empirical and conceptual aspects, the study demonstrates an original, interesting, and most importantly interdisciplinary approach to the topic of color perception and constancy, making an effort to combine psychological research with philosophical enactive theories.

The main objective of this commentary is to discuss what the sensorimotor account of perceptual consciousness could learn from in-

vestigations into phenomenal adaptation in atypical visual conditions such as color rotated spectrum and synesthesia. In the first section, after pinpointing the conceptual and methodological difficulties involved in defining and testing phenomenal adaptation in Grush et al.'s study, I shall deepen our understanding of phenomenal adaptation and analyze various possible readings of this phenomenon. Such readings depend on different interpretations of the contents that are admissible to perceptual consciousness (cf. Hawley & Macpherson 2011). In the second section, the relationship between the color rotation study and the enactive account of color vision is examined in order to demonstrate what consequences the sensorimotor theory may expect from results that confirm it in some respects, but not others. Finally, in the last two sections, I claim that the lessons for enactive theories of color perception may be expanded beyond the implications of the color rotation study. This is verified by looking at confirmatory and challenging cases provided by atypical perceptual conditions and color modifications such as synesthetic color experiences.

2 Phenomenal adaption

The notion “adaptation”, being central to the target article, while used comparably to analogous work on perceptual effects of systematic alteration of sensory input, does not obviously correspond to the unambiguous physiological notion of adaptation, i.e., a decrease over time in the responsiveness of sensory receptors to changed, constantly applied environmental conditions (e.g., Held 1965; Noguchi et al. 2004; Smithson 2005). Distinguishing semantic adaptation (a remapping of color terms and building immediate semantic connections to their proper object referents) from phenomenal adaption, Grush et al. focus on phenomenal aspects of regaining both stimulus constancy and original color arrangement in spite of changes in input. Given that adaptation to numerous alterations in visual input has already been reported in various studies (Kohler 1962, 1963; Anstis 1992; Heuer & Hegele 2008), Grush

and colleagues also hypothesize the possibility of some form of adaptation to a version of the color-inverted spectrum. They designed a series of experiments to assess phenomenal adaptation of visual experience under color rotation by 120°, which leads to “tomatoes [...] causing red qualia again, even if the subject is wearing the rotation gear” (Grush et al. [this collection](#)). The definition of “phenomenal adaptation” in the target article is “a return to normalcy” and “a gaining of color constancy under rotation”. Phenomenal adaptation then, can be understood as the regaining of phenomenal qualities of the pre-rotated color experience while using the rotation equipment for some time, such that experienced colors are stable, constant, and non-rotated, i.e., just like in normal color vision under standard conditions. But there are reasons to think that adaptation is not necessarily a phenomenally-conscious phenomenon. Such an assumption is supported by research with blindsight patients exhibiting, in their unconscious perception, spectral wavelength sensitivity and several other features of color vision adaptation (Stoerig & Cowey 1989, 1991). In addition, adaptation should not be confused with habituation, which is an attentional phenomenon over which subjects reveal some conscious control (Webster 2012). This could help to explain some of the difficulties Grush et al. encountered when trying to prove the occurrence of phenomenal adaptation under color rotation—which are described below.

2.1 Conceptual problems — Defining phenomenal adaptation

For most of the target paper, Grush et al. treat “phenomenal adaptation” as if it has a single, obvious, and straightforward meaning. Only at the end do they briefly hint at different readings of such adaptation depending on the understanding of the notions “phenomenal” or “qualia”. That is, their investigation is driven by certain implicit assumption of phenomenal qualities. However, these terms are quite controversial in philosophy and one may wonder what actually is examined in the study.

Philosophers denying the existence of phenomenal qualities or qualia (e.g., Churchland

1985, 1989; Tye 1995, 2000; Dennett 1988) may understand them in a specific and narrow sense, either as consciously-accessible properties of non-physical mind-dependent phenomenal objects, mental images called sense data (Lewis 1929; Robinson 1994), or intrinsic non-representational properties (Block 1990; Peacocke 1983), or non-physical, ineffable properties of experiences given to their subjects incorrigibly (Dennett 1988, 1991). Nonetheless, qualia may be endorsed in a broader sense, namely as phenomenal character. This use of the term generally refers to introspectively accessible qualitative aspects of one’s mental life, and it is hard to deny that these exist. Phenomenal character of an experience is “what it is like” for a subject to undergo the experience (Shoemaker 1994, 2001; Chalmers 1996; Nagel 1974). While engaging in introspection and focusing attention on the phenomenal character of experience, one is aware of and gets access to certain phenomenal qualities that make up the overall phenomenal character of the experience.

Since there is no single definition of the term “phenomenal qualities” or “qualia”, there might be also more than one reading of the notion of “phenomenal adaptation”. Depending on the particular understanding of phenomenality and the sort of mental states that can have phenomenal qualities or enter phenomenal consciousness, there seem to be different ways of interpreting the phenomenon of adaptation and thus the possibility of different kinds of adaptation. A related matter discussed in the philosophy of perception—between those supporting phenomenal liberalism and those who propose phenomenal conservatism (expansive and restrictive conceptions of the domain of phenomenal consciousness)—is whether there are high-level properties in the content of perception and whether cognitive states have a distinctive and proprietary phenomenology (Bayne 2009; Prinz 2012).

Phenomenal conservatism (e.g., Tye 1995, 2000; Carruthers 2005; Braddon-Mitchell & Jackson 2007; Nelkin 1989), proposing austere perceptual phenomenology, i.e., that the contents of perception are exclusively of the sensory sort, promotes sensory adaptation. Phenom-

enal adaptation, understood as sensory adaptation, has been described in the literature as adaptation to various distortions and systematic alterations of sensory input employing single or multiple modalities. For example, subjects wearing prismatic goggles or lenses inverting a visual scene in terms of color and spatial arrangement can adapt in the course of time to these new settings and become able to act normally, because they develop new visuo-tactile contingencies that allow them to get around and efficiently see and reach for objects (Held 1965). Importantly, such an adaptation directly affects perceptual experience and cannot be explained by correcting judgments. But this may not be the whole story about phenomenal adaptation, since phenomenal character does not have to be limited to sensory experiences, although traditionally it is said to be. In line with phenomenal liberalism, contents of perception may contain high-level properties such as kind properties (e.g., the property of being a tiger; recognizing that something belongs to a certain kind—seeing a tree as a pine tree; Siegel 2006; Bayne 2009), but also causal (the property of one thing's causing another; Strawson 1985; Siegel 2006; Butterfill 2009), and generic properties (the property of being nonspecific; Block 2008; Grush 2007). These properties are abstract, generalized, and cognitive in their nature, yet they can enter into phenomenal contents. Consequently, a liberal conception, allowing cognitive states to possess phenomenal qualities, and phenomenal character to be ascribed to conceptual contents, endorses cognitive phenomenology and thus would opt for phenomenal adaptation in the cognitive aspects of experience.

The debate surrounding cognitive phenomenology involves many different versions and strengths of the claim that the domain of phenomenology extends beyond the sensory (Strawson 1994; Siewert 1998; Pitt 2004; Bayne & Montague 2011; Horgan & Tienson 2002; Kriegel 2002, 2007). Irrespective of its particular varieties, such a view raises alternative interpretations of Grush et al.'s results. It suggests that phenomenal adaptation may be present not only in sensory but also in cognitive aspects of experience. Both perceptual and cognitive states

determine how we experience the world and adapt to changes in our surroundings, because they both exhibit their own phenomenal characters—something it is like to be in such a state for the subject (Chalmers 1996, p. 10; Strawson 1994; Montague & Bayne 2011).

Moreover, conceptual contents seem able to modify the phenomenal character of perceptual states; they can cognitively penetrate our perception (Raftopoulos 2005; Macpherson 2012; Siegel 2012). An interdisciplinary approach to the cognitive penetrability of perception assumes that there are various ways in which conscious perception can be affected by cognition—i.e., by thoughts, beliefs, desires, judgments, intentions, moods, emotions, expectations, knowledge, previous experiences, and memories (Frith & Dolan 1997; Bar 2003; McCauley & Henrich 2006; McCauley & Henrich 2006; Raftopoulos 2009; Vuilleumier & Driver 2007; Stokes 2012; Deroy 2013; Wu 2013; Vetter & Newen 2014; Briscoe 2014; Nanay 2014; Lupyan 2015). In other words, higher cognitive states not only have causal influence on the contents of perception, they are also explanatorily relevant in accounting for the processing of perceptual systems. It has been shown that semantic contents and categories play a critical role in perception, even in early sensory processing (cf. Mroczko et al. 2009; Mroczko-Wąsowicz & Nikolić 2013). This may be exemplified in the connection between language and color vision. For example, languages with a larger number of generic color terms such as Russian have an impact on color perception (Winawer et al. 2007).

Such an integrative cogno-sensory approach, combining high-level cognitive and low-level sensory aspects, is also manifested in recent theories of concepts relating the possession of concepts to perceptual adaptation in various ways (Machery 2009; Prinz 2010; Noë this collection). For instance, being sensitive and showing a discriminative response to certain kinds of objects or combinations of features corresponds to having concepts for the related kinds of objects (Machery 2009; Deroy 2013, 2014). According to the ability-based account of conceptuality, one can reveal skillful understanding of

concepts in a perceptual, practical, or emotional way, meaning that the possession of concepts is a condition that informs and is informed by our able engagement with things (Noë 2012, [this collection](#); cf. Wittgenstein 1953). This indicates a close interdependence between conceptuality and sensorimotor processing.

Consequently, phenomenal adaptation, in light of the enactive theory, would mean enactive adaptation and learning a new set of skills in the form of new sensorimotor contingencies and related dependencies, such as behavioral dispositions, predictive possibilities, and cognitive, aesthetic, and emotional reactions. Enactive adaptation would entail an application of sensorimotor skills to conceptual understanding, and as such it could be seen as adaptation in the cognitive aspects of experience with altered expectations or beliefs about or sensitivity to kinds of objects encountered in perceptual experience. This phenomenally liberal reading would provide an appropriately more capacious notion than the adaptation of the pure sensory sort offered by Grush et al.

To sum up, departing from a distinction between phenomenal conservatism that accepts perceptual phenomenology with solely sensory contents and phenomenal liberalism that acknowledges higher-level contents of perception and cognitive phenomenology (Bayne 2009; Montague & Bayne 2011), I differentiate between adaptation of the purely sensory sort and adaptation in the cognitive aspects of experience. The distinction is used to show the contrast in understandings of the notion of “phenomenal adaptation” between the target article and this commentary. Grush et al. seem to suggest that phenomenal and (non-phenomenal) semantic adaptation are different forms of a more general phenomenon of adaptation. However, they do not give any explicit example of the genus of adaptation of which these later are a species. I contend, in turn, that there is no need to produce such subclasses of the notion; semantic adaptation involving higher-level non-sensory states may also be understood as phenomenal. Thus, the reading of adaptation I put forward pertains jointly to the phenomenal and semantic aspects of regaining of stimulus

constancy; it assumes a recovery of prototypical color-object associations both in phenomenal experience and in semantic reference in spite of changes in input. This follows from phenomenal liberalism.

The proposed view is that being processed in the headway of phenomenal adaptation is phenomenal character, understood in an expansive liberal way that includes high-level contents. Therefore phenomenal adaptation is considered to be the adjustment of cognitive aspects of experience.

2.2 Methodological problems

Dissociating semantic from phenomenal adaptation is problematic. This is because they are interconnected. It is hard to think about the occurrence of semantic adaptation without phenomenal adaptation taking place and vice versa – semantic adaptation is methodologically necessary for detecting phenomenal adaptation. This presumed correlation might be the reason why, when faced with difficulties finding phenomenal adaptation to a color-rotated scene, the investigators could not confirm any reliable Stroop results for semantic adaptation. In addition, it should be noted that Stroop-type tasks contain two components of competition—semantic and perceptual (Stroop 1935; Nikolić et al. 2007; Mroczko et al. 2009)—and as such they exhibit limitations in differentiating between semantic and perceptual aspects of the phenomena tested.

To assess the occurrence of phenomenal color adaptation under rotation, that is, the process of the normal phenomenal appearance of objects returning, Grush et al. used the memory color effect (Hansen et al. 2006), aesthetic judgments of food and people, and subjective reports from their test persons.

It has often been assumed that subjective introspective reports are a generally reliable mode of first-person access to one’s current conscious states or processes (Descartes 1984; Locke 1689/1979; Hume 1978; Brentano 1973; Husserl 1982; Chalmers 2003; Gertler 2001; Horgan et al. 2006; Horgan & Kriegel 2007; Varela 1996; Rees & Frith 2007; Hurlburt &

Schwitzgebel 2007; Hohwy 2011). However, this assumption is also problematic. Arguments for introspective scepticism or even criticism of introspective methodology pose genuine threats to the trustworthiness of this approach (see Bayne [this collection](#), for a discussion of such views). Because of this ambivalence one needs to be careful when using subjective reports as a source of or support for the results presented.

Certain doubts about whether subjective reports are trustworthy enough come from the fact that introspection delivers solely first-person, unverifiable, private data, and thus it is unscientific and often fallible (Dennett 1991; cf. Zmigrod & Hommel 2011). In addition, subjects tested are often uncertain or disagree about what the introspective access actually provides (Bayne & Spener 2010; see also Bayne [this collection](#)) and have difficulty describing their own conscious experiences (Schwitzgebel 2008). Nonetheless, this is not to deny that they have *some* first-person knowledge of phenomenal consciousness.

The specific reasons one may have for doubting the findings in the context of Grush et al.'s study are related to the fact that investigators were also the test subjects. We should avoid involving persons who know the hypothesis when conducting the experiments, who in this case tested and evaluated themselves at the same time. Knowing the research question and the expected or desired results may bias any study.

3 Implications of the color rotation study for sensorimotor enactivism

Grush et al.'s work could be extended to manifest a broader range of philosophical implications than those they have mentioned; but, as the authors state at the end of their article, this has been left for future philosophical and psychological investigation. Referring to a general theoretical framework of perception such as the enactive approach, Grush and colleagues apply the lessons of their study to sensorimotor enactivism of perception without considering other options such as ecological and active perception approaches as potential targets (Gibson 1979;

Ballard 1991; Mossio & Taraborelli 2008; Taraborelli & Mossio 2008). However, since their hypothesis focuses on the nature of perception based on couplings between sensory stimulation and motor activity, it appears justified to focus on the sensorimotor version of the enactive account, which emphasizes an active exploration of the environment determining in this way the content and modality of conscious experience (O'Regan & Noë 2001; Noë 2005).

Sensorimotor theory has been supported by research on sensory substitution (Proulx & Störig 2006) and adaptation in haptic perception, as observed in mirror therapy for phantom limb pain and in the rubber hand illusion (Ramachandran & Rogers-Ramachandran 1996; Botvinick & Cohen 1998). Most relevantly, supporting evidence for the sensorimotor theory of color perception was found in a study on adaptation to half-split colored goggles (left-field blue/right-field yellow), which introduced an artificial contingency between eye movements and color changes (Bompas & O'Regan 2006b; cf. Kohler 1962). These results have left the possibility of similar sensorimotor adaptation to any arbitrarily-chosen colors open. According to the account of enactive vision, sensorimotor principles are fully capable of explaining adaptation to alterations in spatial or color-relevant features of input (Noë 2005). The adaptation can be achieved by resuming constancy through learning a new set of sensorimotor contingencies, i.e., patterns of dependence between sensory stimulation and movements, corresponding to new features of the input. Understanding these dependencies provides the required sensorimotor knowledge that enables perceptual experience.

The experimental protocol of Grush et al.'s study directly refers to an enactive account of color (O'Regan & Noë 2001; Noë 2005; Bompas & O'Regan 2006a, 2006b). The authors' hypothesis regarding color constancy and phenomenal color adaptation under color rotation is compatible with predictions made by sensorimotor enactivism; the induced adaptation to a remapped spectrum was supposed to imitate a naturally-occurring process of learning sensor-

imotor contingencies. The results obtained in this pilot study, although not entirely usable and interpretable, may yet provide food for thought to enactive theory, since they offer some interesting insights into supportive evidence and the difficulties that the theory needs to integrate and deal with.

3.1 An enactivist explanation of the results

Subjective reports concerning adaptation to color constancy, understood as achieving stability of color experience irrespective of visual conditions, confirm what the enactive theory would expect. This means that when switching between standard visual conditions and color rotation, and at the same time being active in the color environment through altering color-critical conditions such as illumination, viewing angles, or movements, the test persons exhibited temporary disruption of color constancy leading to an immediate change of perceived hues. This is allegedly due to the change of sensorimotor contingencies involved in this experience.

However, when a new set of sensorimotor regularities becomes established, color constancy is resumed, so that the subjects gain the capacity for color constancy under rotation and then come back to normal color constancy when having non-rotated visual input, i.e., the colors that are stable are different in the two conditions. Hence, after a period of time for learning new dependencies, color constancy is restored and the mentioned modifications of visual conditions, such as lighting, have no effect on the phenomenal character of color experience.

An interesting observation and an important point for further deliberation on the development of phenomenal color adaptation is delivered in the subjective report of one of the test subjects, who at the end of his six-day color rotation period suddenly began to be confused about whether his visual input was still rotated or not, because everything appeared normal. Since he ceased to feel a sense of novelty and strangeness, he was not sure if he was in a situation of (1) normal color vision, or rather (2) adaptation to color constancy under rotation—at least until he expli-

citly reflected on the colors of the surrounding objects. Although he was evidently in state (2), thus experiencing stability of rotated colors, one may suppose that his confusion about which colors were ‘normal’ in which condition might also indicate the time in which subjects could begin to develop an ability amounting to (3) phenomenal color adaptation under rotation with colors akin to genuine colors in situation (1). Speculations envisioning the occurrence of this adaptation after a longer period than the duration of the current test do not seem completely unjustified. What would be needed here are further studies that not only cover a longer time frame of color rotation, but also focus on searching for a characteristic marker signaling when, within a very smooth transition between (2) and (3), phenomenal adaptation under rotation (stage(3)) actually begins. This would be similar to “the feeling of novelty/strangeness”-marker within the transition between (1) and (2), signaling color rotation. The lack of this marker and the occurrence of the feeling of normality would indicate that color constancy under rotation has arisen.

The memory color effect was used by Grush et al. as a method of assessment for phenomenal color adaptation under rotation. It is an effect of processing colors of objects with typical colors that affects the experience of pairings of colors and shapes (Hansen & Gegenfurtner 2006; Hansen et al. 2006). The authors explain the effect by top-down influences of expectations. But it may also be explained by, for example, cognitive penetration of color experience by beliefs (Macpherson 2012) or sensory adaptation through exposure manifesting itself by responding differently to various kinds of objects or co-occurring features (e.g., arrangements of objects’ shapes and their typical colors; Deroy 2013, 2014). All these descriptions express some aspect of the phenomenal liberalism discussed earlier, and as such they seem more or less equally plausible for supporting the proposed reading of phenomenal color adaptation under rotation as adaptation in the cognitive aspects of experience. In standard visual conditions, the memory color effect may suggest that expectations or beliefs about a proper color for a certain kind of objects exert top-down influence on the actual color pro-

cessing of these objects, their shapes, etc. Thus, the lessened magnitude of the memory color effect under color rotation, as found in the study, shows that the associations of objects with their prototypical colors become weaker and may even get replaced by other associations with new prototypical colors.

This outcome is interestingly combined by Grush et al. with the aforementioned confusion stage (between (1) and (2)) acquired at the end of the color rotation period, when the subject stops having the feeling of novelty and therefore confuses his rotated color experiences with the normalcy felt when perceiving in standard visual conditions. Both of these results imply not only a decreasing strength of the old prototypical color associations, but also the emergence of new associations. Such an emergent set of dependencies is clearly compatible with enactive predictions. The adaptation that took place due to color rotation and that has been demonstrated by the memory color effect appears to be general. This means it is not just a matter of specific associations of colors with particular objects seen during rotation. The adaptation refers to the perceptual system as a whole and its expectations, beliefs, or sensitivity, contributing to a discriminative response to kinds of objects in general. For example, the adaptation might manifest itself as the regaining of a grasp of the way things are colored, as altered cognitive states (cognitive aspects of experience) about what red things generally look like or what red is like.

3.2 Problems with a definitive confirmation of enactivist ideas

Obviously the study protocol would have been more plausible if color constancy had been tested in a controlled way with a relevant objective method and not only confirmed by first-person reports. For example, brain imaging techniques would be suitable for detecting temporary changes in perceptual states. Also, comparing the effect with a proper control group, matching the test group for gender, age, and color-related experience (e.g., education, profession), would certainly increase the strength of the findings, providing more evidence for sen-

sorimotor adaptation to color constancy. Because transformations in qualitative experience may be explained in terms of a dynamic model of interdependence between sensory inputs and embodied activity (Hurley & Noë 2003), phenomenal differences between color experiences can be accounted for by different actions. Therefore to exclude the sensorial interpretation, the control group would not be actively exploring their color environment, would not change the rotated visual input through their own actions, and thus according to the enactive theory would not develop new sensorimotor dependencies allowing stable color perception.

For genuine phenomenal color adaptation different results were observed, i.e., the regaining of non-rotated color constancy while using the rotation equipment was not successfully established—subjective reports and objective assessments made with the memory color effect and aesthetic judgments of color-rotated food and people have shown that subjects only started to adapt in late-rotation, at the end of the possible adaptation period. Difficulties in robustly confirming phenomenal color adaptation under rotation are certainly not encouraging news for the enactive view of color. They could even be interpreted as a falsification of this theory. However, according to the investigators, this is still not decisive, and they speculate that the reason for this unfavorable outcome could be the lack of time allowed for relearning the relevant sensorimotor regularities. Indeed, for someone whose phenomenal color qualities remained rotated and did not revert to the genuine color phenomenology, i.e., for whom tomatoes continued to look blue, but did not reappear as red, this may be the case, because perceptual learning, here resulting in action-sensation coupling, is a relatively slow process and its timing varies from one individual to another (Goldstone 1998; Seitz & Watanabe 2005). Such an explanation remains in line with the sensorimotor account of perception and cannot be excluded without further studies. On the other hand, it may be also possible that the development of adaptation under rotation took place unconsciously and therefore was not reported by the subjects.

4 Atypical color conditions in synesthesia

The lessons for enactive theories of color perception, pointed to by Grush et al. in their target article, may also be expanded by including challenges constituted by other atypical color conditions, namely synesthetic color experiences.

Synesthesia is traditionally considered to be a phenomenon in which the stimulation of one sensory or cognitive pathway (the inducer) elicits involuntary and consistent sensory experiences (the concurrent) in the same or another modality (Baron-Cohen et al. 1987; Baron-Cohen & Harrison 1997; Ramachandran & Hubbard 2001a, 2001b). As a result, the stimuli corresponding to the inducer and the experiences associated with the concurrent form a highly integrated percept—a phenomenally-unified experience which may cover not only sensory modalities, but also various mental domains including conceptual, emotional, bodily, and motor aspects (Mroczko-Wąsowicz & Werning 2012; Mroczko-Wąsowicz 2013). Such unification incorporates the central system and early stages of processing. Some synesthetes see colors when dealing with letters or numerals. Individuals with another kind of synesthesia perceive colored patterns in space when hearing sounds. The prevalence of the phenomenon depends on the particular type of synesthetic association, with grapheme-color synesthesia being the most common (Cytowic & Wood 1982; Mroczko-Wąsowicz & Nikolić 2013).

Color sensations are the most frequent synesthetic concurrents (Marks & Odgaard 2005), demonstrating color opponent properties and neural representations more or less similar to veridical color experiences (Nikolić et al. 2007; Hubbard et al. 2005; cf. Hupé et al. 2012; Van Leeuwen et al. 2010). For some forms of synesthesia color concurrents may also originate from information processing in regions of the cortex other than the visual. Recent neuroimaging studies demonstrate that synesthetic colors for numbers or mathematical formulas may also be produced when the visual cortex is not involved, i.e., by the activation of temporal, parietal, and frontal brain areas (Bor et al. 2007;

Hupé et al. 2012; Brogaard et al. 2013). This suggests that information processing in non-visual brain regions may be a source of concurrent colors and therefore some forms of synesthesia can be seen as high-level perception proceeding via non-standard mechanisms. Such high-level synesthetic color perception for mathematical skills, though quite unusual, may provide supportive evidence for the conception of phenomenal liberalism and cognitive phenomenology.

5 Synesthetic colors and sensorimotor enactivism

Given that synesthesia, similarly to the color rotation gear, involves systematic distortions of color perception that are consciously experienced by the subjects, analyzing synesthetic experiences appears relevant in the context of the present discussion. This is why proponents of the sensorimotor theory of color perception might be interested in examining whether their postulates also apply in cases employing such synesthetic color-addition gear (cf. Hurley & Noë 2003; Fingerhut 2011; Mroczko-Wąsowicz & Werning 2012; Seth 2014; Ward 2012). The relevant propositions of enactivism may be stated as follows:

- a. determining the modality of perceptual experience by specific sensorimotor signature (i.e., dependency between sensory stimulations and the activity of the perceiver, including their motor actions, bodily changes, or behavioral skills), as well as a necessary possession of such sensorimotor knowledge of contingencies enabling any perception;
- b. flexibility of perceptual experience manifested in the ease of its modification and adaptation based on learning a new set of sensorimotor contingencies (Noë 2005);
- c. and finally epistemic reliability of conscious perceptual experiences and their counterfactual richness (Metzinger 2014; Seth 2014; see also Seth this collection).

The above basic assumptions underlying sensorimotor enactivism of perception may be challenged by synesthesia in the following way:

- A. Synesthetic concurrent percepts (e.g., visual experiences) are generated internally, not via a direct relation of a synesthete with the surrounding environment. They are triggered without employing the regular sensorimotor signature related to these concurrents, like eye saccades in normal vision. For such permanent inducer-concurrent couplings, the concurrent modality and its experiences are never related to their normal sensorimotor signature.
- B. Synesthetic associations cannot be learned or adapted to, in contrast to various manipulations of sensory input such as, for example, spatial displacement, color inversion, or auditory-visual sensory substitution (sometimes called an artificial synesthesia), which are used by sensorimotor enactivists as examples of the perceptual system's adaptation involving an appropriate adjustment of sensorimotor contingencies. Unlike the majority of learned pairings, synesthetic associations are rigid and not flexible enough to adapt, irrespective of the amount of exposure to contradictory experiences or training (Baron-Cohen et al. 1993; Deroy & Spence 2013).
- C. As a final point, although synesthetic colors are reported to be as vivid as non-synesthetic colors, synesthetes immediately detect the difference between them, which confirms the absence of perceptual presence or phenomenal transparency in synesthesia, meaning its opacity or experiential unrealness, which is the availability of earlier processing stages to attention (Metzinger 2003a, 2003b, 2014; Seth 2014).

As a kind of reply to these challenges, sensorimotor enactivists could claim that enactivism focuses on standard perceptual mechanisms and therefore has difficulties explaining perception-

like experiences in synesthesia, as well as that synesthetic concurrents (often colors) lack some important features of typical perceptual experiences and properties of sensorimotor engagement, e.g., corporeality. However, this would not really be explanatory. One possible way of vindicating how the enactive theory could accommodate such atypical non-adaptive color conditions is to claim that there is actually no need for synesthetic colors to adapt, because they do not carry any information about the colors of objects in the synesthetes' environment—whereas adaptation is a retrieval of how things are colored. To put it in enactive terms, synesthetic colors do not figure in patterns of appearance reflecting dynamic relations between perceiver, object, and light (Ward 2012). Unlike the rotation gear, synesthesia does not determine the way things appear to the perceiver; i.e., the way worldly objects and surfaces modify the light is not affected.

6 Discussion

As concluded by Grush et al. the results obtained in their study show that color experiences changed in the early stage of application of the rotation gear and became stable, that is, they adapted to color constancy in the late rotation stage, without however consistently showing significant phenomenal adaptation by the end of the test. The investigators leave open a potential explanation of this outcome. The difference between this result and other studies, in which achieving phenomenal adaptation to spatial displacement or luminance inversion was more successful, may suggest that color sensations are special properties of early visual processing relatively difficult to phenomenally adapt as well as more resistant to penetration and manipulation by cognition (Fodor 1983; Pylyshyn 1999; Brogaard & Gatzia forthcoming; but cf. Macpherson 2012; Siegel 2012; Vetter & Newen 2014). At least this seems to be the case for the general population.

Synesthesia, although not considered to be an adaptive plastic phenomenon, may be a case in which some modifications take place, such as cognitive penetration of perception including

early sensory processing. Synesthetic colors are frequently modified by cognitive operations, conceptual contents, contextual expectations, linguistic modulation, cultural factors, and other semantic knowledge mechanisms (Dixon et al. 2000; Simner 2007, 2012; Meier 2013; Mroczko et al. 2009; Mroczko-Wąsowicz & Nikolić 2013, 2014). Since synesthetes are able to penetrate this early aspect of vision it would be interesting to investigate whether synesthetically-perceived colors change under rotation. If so, is this in the same or in a different way to non-synesthetic, phenomenally-transparent colors?

Synesthetic colors are in some respects similar to the color experiences of rotation-gear wearers. In both cases, subjects are aware of the fact that what they see is not reliably colored, i.e., that their abnormal color experiences are not actual colors of the surroundings.

On the other hand, these color experiences differ remarkably from each other. Whereas rotation gear wearers' color experiences are able to adapt to fall in line with what the subjects know to be true about colors of the things around them, synesthetes' color experiences do not display such flexibility. Thus, a theoretically founded hypothesis is that irrespective of the form of color synesthesia to be used in a color rotation experiment (e.g., grapheme-color, sound-color, time unit-color synesthesia) synesthetic colors would not alter.

Admittedly, the sensorimotor theory of color has difficulties explaining many of the features of the phenomenon of synesthesia, but this does not mean it is completely useless in the context of synesthesia. The theory could be used to account for the asymmetry in adaptation capability between those experiencing synesthetic and non-synesthetic colors. From the perspective of the enactive view of color, it could be proposed that the rotation gear interferes with regular color perception, because the equipment introduces a new set of sensorimotor dependencies. This is the reason why after wearing the rotation gear for some time and acclimatizing to the conditions, the rotated colors begin to appear normal, that is, the subjects' ability to perceive original colors returns

—such that phenomenal color adaptation under rotation takes place. Unlike the rotated color perception of non-synesthetes, synesthetic subjects (associators) do not experience their additional colors as attributed to perceived objects but as seen in their “mind’s eye”. Therefore the concurrent colors do not affect their ability for regular color vision. An explanation of why there is no phenomenal color adaptation in synesthesia could be that synesthetic colors just fail to adapt because they do not need to make room for non-synesthetic colors (cf. Ward 2012). Thus, in line with the sensorimotor account, we could interpret the difference between ordinary color perception and color synesthesia as a difference between real and seeming engagement. However, another type of synesthetes, projectors, who see colors as projected onto inducing objects, may require a different explanation. Since this group of synesthetes demonstrates an external frame of reference for their synesthetic colors, projectors' phenomenal color adaptation under rotation might be comparable to the adaptation of non-synesthetes. A testable empirical prediction here would be that it is possible for projector synesthetes' colors to adapt after using a rotation equipment specially adjusted to interfere with internally-generated concurrent color experiences. Depending on the outcome of this prospective study, which could be designed in such a way that it would take into account all the differences related to color phenomenology, the sensorimotor theory of color may gain new insights into the threat of synesthesia.

What sensorimotor enactivism can also learn from studies on phenomenal adaptation in various perceptual conditions is that the notion of “phenomenal adaptation” applies to some conditions, but not others. There may also be different senses of the notion, and apart from “adaptation of the sensory sort” a rigorous analysis should consider “adaptation in the cognitive aspects of experience”—an expansive interpretation supported by phenomenal liberalism and cognitive phenomenology. In addition, the extent to which the phenomenon of adaptation may develop varies among various perceptual conditions. No matter the exact magnitude of

the adaptive effects discussed, the very existence of phenomenal adaptation to alterations of sensory input, or its general lack, needs to be fully integrated by philosophical theories, especially by sensorimotor enactivist theories of perception that attempt to account for all the dynamics related to perception. This adaptation highlights that perceptual experience is more flexible and variable than usually presumed.

Acknowledgements

I would like to thank the editors and anonymous reviewers for very helpful comments on earlier versions of this commentary.

References

- Anstis, S. (1992). Visual adaptation to a negative, brightness-reversed world: Some preliminary observations. In G. Carpenter & S. Grossberg (Eds.) *Neural networks for vision and image processing* (pp. 1-14). Cambridge, MA: MIT Press.
- Ballard, D. H. (1991). Animate vision. *Artificial Intelligence*, 48 (1), 57-86. [10.1016/0004-3702\(91\)90080-4](https://doi.org/10.1016/0004-3702(91)90080-4)
- Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object recognition. *Journal of Cognitive Neuroscience*, 15 (4), 600-609. [10.1162/089892903321662976](https://doi.org/10.1162/089892903321662976)
- Baron-Cohen, S., Wyke, M. & Binnie, C. (1987). Hearing words and seeing colours: An experimental investigation of a case of synaesthesia. *Perception*, 16 (6), 761-767. [10.1068/p160761](https://doi.org/10.1068/p160761)
- Baron-Cohen, S., Harrison, J., Goldstein, L. H. & Wyke, M. (1993). Coloured speech perception: Is synaesthesia what happens when modularity breaks down? *Perception*, 22 (4), 419-426. [10.1068/p220419](https://doi.org/10.1068/p220419)
- Baron-Cohen, S. & Harrison, J. E. (Eds.) (1997). *Synaesthesia: Classic and contemporary readings*. Cambridge, MA: Blackwell.
- Bayne, T. (2009). Perception and the reach of phenomenal content. *The Philosophical Quarterly*, 59 (236), 385-404. [10.1111/j.1467-9213.2009.631.x](https://doi.org/10.1111/j.1467-9213.2009.631.x)
- (2015). Introspective insecurity. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Bayne, T. & Montague, M. (Eds.) (2011). *Cognitive Phenomenology*. New York, NY: Oxford University Press.
- Bayne, T. & Spener, M. (2010). Introspective humility. *Philosophical Issues*, 20 (1), 1-22. [10.1111/j.1533-6077.2010.00176.x](https://doi.org/10.1111/j.1533-6077.2010.00176.x)
- Belmore, S. C. & Shevell, S. K. (2011). Very-long-term and short-term chromatic adaptation: Are their influences cumulative? *Vision Research*, 51 (3), 362-366. [10.1016/j.visres.2010.11.011](https://doi.org/10.1016/j.visres.2010.11.011)
- Block, N. (1990). Inverted earth. *Philosophical Perspectives*, 4, 53-79. [10.2307/2214187](https://doi.org/10.2307/2214187)
- (2008). Consciousness and cognitive access. *Proceedings of the Aristotelian Society*, 108, 289-317.
- Bompas, A. & O'Regan, J. K. (2006a). Evidence for a role of action in color perception. *Perception*, 35 (1), 65-78. [10.1068/p5356](https://doi.org/10.1068/p5356)
- (2006b). More evidence for sensorimotor adaptation in color perception. *Journal of Vision*, 6 (2), 142-153. [10.1167/6.2.5](https://doi.org/10.1167/6.2.5)

- Bor, D., Billington, J. & Baron-Cohen, S. (2007). Savant memory for digits in a case of synaesthesia and Asperger syndrome is related to hyperactivity in the lateral prefrontal cortex. *Neurocase*, 13, 311-319.
[10.1080/13554790701844945](https://doi.org/10.1080/13554790701844945)
- Botvinick, M. & Cohen, J. (1998). Rubber hands “feel” touch that eyes see. *Nature*, 391 (756). [10.1038/35784](https://doi.org/10.1038/35784)
- Braddon-Mitchell, D. & Jackson, F. (2007). *Philosophy of mind and cognition*. Oxford, UK: Blackwell.
- Brentano, F. (1973). Embodied Prediction. In A. C. Rancurello, D. B. Terrell & L. L. McAlister (Eds.) (Trans.) London, UK: Routledge.
- Briscoe, R. E. (2014). Do intentions for action penetrate visual experience? *Frontiers in Psychology*, 5 (1265).
[10.3389/fpsyg.2014.01265](https://doi.org/10.3389/fpsyg.2014.01265)
- Brookes, J. (1992). The autonomy of color. In K. Lennon & D. Charles (Eds.) *Reduction, explanation and realism* (pp. 421-465). Oxford, UK: Oxford University Press.
- Brogaard, B., Vanni, S. & Silvano, J. (2013). Seeing mathematics: Perceptual experience and brain activity in acquired synesthesia. *Neurocase*, 19, 566-575.
[10.1080/13554794.2012.701646](https://doi.org/10.1080/13554794.2012.701646)
- Brogaard, B. & Gatzia, D. E. (forthcoming). *Is color experience cognitively penetrable? Topics in Cognitive Science: Special Issue on Cortical Color*.
- Burge, T. (2010). *Origins of objectivity*. New York, NY: Oxford University Press.
- Butterfill, S. (2009). Seeing causes and hearing gestures. *The Philosophical Quarterly*, 59 (236), 405-428.
[10.1111/j.1467-9213.2008.585.x](https://doi.org/10.1111/j.1467-9213.2008.585.x)
- Byrne, A. & Hilbert, D. (Eds.) (1997). *Readings on color, volume 1: The philosophy of color, Consciousness: Essays from a higher-order perspective*. Oxford, UK: Oxford University Press.
- Carruthers, P. (2005). Conscious experience versus conscious thought. In P. Carruthers (Ed.) *Consciousness: Essays from a Higher-Order Perspective*. Oxford, UK: Oxford University Press.
- Casati, R. (1990). What is wrong in inverting spectra? *Teoria*, 10, 183-6.
- Chalmers, D. (1996). *The conscious mind: In search of a fundamental theory*. New York, NY: Oxford University Press.
- (2003). The content and epistemology of phenomenal belief. In Q. Smith & A. Jokic (Eds.) *Consciousness: New philosophical perspectives* (pp. 220-272). Oxford, UK: Oxford.
- Churchland, P. (1985). Reduction, qualia, and direct introspection of brain states. *Journal of Philosophy*, 82 (1), 8-28. [10.2307/2026509](https://doi.org/10.2307/2026509)
- (1989). Knowing qualia: A reply to Jackson. In Y. Nagasawa, P. Ludlow & D. Stoljar (Eds.) *A neurocomputational perspective* (pp. 163-178). Cambridge, MA: MIT Press.
- (2005). Chimerical colors: Some phenomenological predictions from cognitive neuroscience. *Philosophical Psychology*, 18 (5), 527-560.
[10.1080/09515080500264115](https://doi.org/10.1080/09515080500264115)
- Clark, A. (1985). Spectrum inversion and the color solid. *Southern Journal of Philosophy*, 23 (4), 431-443.
[10.1111/j.2041-6962.1985.tb00413.x](https://doi.org/10.1111/j.2041-6962.1985.tb00413.x)
- Cohen, J. (2001). Color, content, and Fred: On a proposed reductio of the inverted spectrum hypothesis. *Philosophical Studies*, 103 (2), 121-144.
- Cohen, J. D. & Matthen, M. (Eds.) (2010). *Color ontology and color science*. Cambridge, MA: MIT Press.
- Cytowic, R. E. & Wood, F. B. (1982). Synesthesia: I. A review of major theories and their brain basis. *Brain Cognition*, 1 (1), 23-35.
[10.1016/0278-2626\(82\)90004-5](https://doi.org/10.1016/0278-2626(82)90004-5)
- Dennett, D. (1988). Quining qualia. In A. J. Marcel & E. Bisiach (Eds.) *Consciousness in modern science* (pp. 42-77). Oxford, UK: Oxford University Press.
- (1991). *Consciousness explained*. Boston, MA: Brown and Little.
- Deroy, O. (2013). Object-sensitivity versus cognitive penetrability of perception. *Philosophical Studies*, 162 (1), 87-107.
[10.1007/s11098-012-9989-1](https://doi.org/10.1007/s11098-012-9989-1)
- (2014). Modularity of perception. In M. Matthen (Ed.) *Oxford Handbook of Philosophy of Perception*. Oxford, UK: Oxford University Press.
[10.1093/oxfordhb/9780199600472.013.028](https://doi.org/10.1093/oxfordhb/9780199600472.013.028)
- Deroy, O. & Spence, C. (2013). Why we are not all synesthetes (not even weakly so). *Psychonomic Bulletin & Review*, 20 (4), 643-664.
[10.3758/s13423-013-0387-2](https://doi.org/10.3758/s13423-013-0387-2)
- Descartes, R. (1984). *Meditations on first philosophy; The philosophical writings of Descartes*. Cambridge, UK: Cambridge University Press.
- Dixon, M. J., Smilek, D., Cudahy, C. & Merikle, P. M. (2000). Five plus two equals yellow. *Nature*, 406 (365).
[10.1038/35019148](https://doi.org/10.1038/35019148)
- Fingerhut, J. (2011). Sensorimotor signature, skill, and synaesthesia. Two challenges for enactive theories of perception. In J. Fingerhut, S. Flach & J. Söffner (Eds.) *Habitus in habitat III. Synaesthesia and kinaesthetics* (pp. 101-120). Berne, GER: Peter Lang.
- Fodor, J. (1983). *The modularity of mind: An essay on faculty psychology*. Cambridge, MA: MIT Press.

- Frith, C. & Dolan, R. J. (1997). Brain mechanisms associated with top-down processes in perception. *Philosophical Transactions of the Royal Society of London: Series B, Biological Sciences*, 352 (1358), 1221-1230. [10.1098/rstb.1997.0104](#)
- Gertler, B. (2001). Introspecting phenomenal states. *Philosophy and Phenomenological Research*, 63 (2), 305-328. [10.2307/3071065](#)
- Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin Harcourt.
- Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology*, 49, 585-612. [10.1146/annurev.-psych.49.1.585](#)
- Grush, R. (2007). A plug for generic phenomenology. *Behavioral and Brain Sciences*, 30 (3), 504-505. [10.1017/S0140525X07002841](#)
- Grush, R., Jaswal, L., Knoepfler, J. & Brovold, A. (2015). Visual Adaptation to a Remapped Spectrum. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Hansen, T., Olkkonen, M., Walter, S. & Gegenfurtner, K. R. (2006). Memory modulates color appearance. *Nature Neuroscience*, 9 (11), 1367-1368. [10.1038/nm1794](#)
- Hansen, T. & Gegenfurtner, K. R. (2006). Color scaling of discs and natural objects at different luminance levels. *Visual Neuroscience*, 23 (3-4), 603-610. [10.1017/S0952523806233121](#)
- Hardin, C. L. (1993). *Color for philosophers*. Indianapolis, IN: Hackett.
- Hawley, K. & Macpherson, F. (Eds.) (2011). *The admissible contents of experience*. Hoboken, NJ: Wiley-Blackwell.
- Held, R. (1965). Plasticity in sensory-motor systems. *Scientific American*, 213 (5), 84-94.
- Heuer, H. & Hegele, M. (2008). Constraints on visuo-motor adaptation depend on the type of visual feedback during practice. *Experimental Brain Research*, 185 (1), 101-110. [10.1007/s00221-007-1135-5](#)
- Heuer, H. & Rapp, K. (2011). Active error corrections enhance adaptation to a visuo-motor rotation. *Experimental Brain Research*, 211 (1), 97-108. [10.1007/s00221-011-2656-5](#)
- Hilbert, D. R. & Kalderon, M. E. (2000). Color and the inverted spectrum. In S. Davis (Ed.) *Vancouver studies in cognitive science*. New York, NY: Oxford University Press.
- Hohwy, J. (2011). Phenomenal variability and introspective reliability. *Mind & Language*, 26 (3), 261-286. [10.1111/j.1468-0017.2011.01418.x](#)
- Horgan, T., Tienson, J. & Graham, G. (2006). Internal-world skepticism and mental self-presentation. In U. Kriegel & K. Williford (Eds.) *Self-representational approaches to consciousness* (pp. 191-207). Cambridge, MA: MIT Press.
- Horgan, T. & Kriegel, U. (2007). Phenomenal epistemology: What is consciousness that we may know it so well? *Philosophical Issues*, 17 (1), 123-144. [10.1111/j.1533-6077.2007.00126.x](#)
- Horgan, T. & Tienson, J. (2002). *Philosophy of mind: Classical and contemporary readings*. Oxford, UK: Oxford University Press.
- Hubbard, E. M., Arman, A. C., Ramachandran, V. S. & Boynton, G. M. (2005). Individual differences among grapheme-colour synaesthetes: Brain-behavior correlations. *Neuron*, 45, 975-985. [10.1016/j.neuron.2005.02.008](#)
- Hume, D. (1978). *A treatise of human nature*. Oxford, UK: Clarendon.
- Hupé, J. M., Bordier, C. & Dojat, M. (2012). The neural bases of grapheme-color synesthesia are not localized in real color-sensitive areas. *Cerebral Cortex*, 22, 1622-1633. [10.1093/cercor/bhr236](#)
- Hurlburt, R. T. & Schwitzgebel, E. (2007). *Describing inner experience? Proponent meets skeptic*. Cambridge, MA: MIT Press.
- Hurley, S. L. (1998). *Consciousness in action*. London, UK: Harvard University Press.
- Hurley, S. & Noë, A. (2003). Neural plasticity and consciousness. *Biology & Philosophy*, 18, 131-168.
- Husserl, E. (1982). Ideas. In T. E. Klein & W. E. Pohl (Eds.) *Book I*. Dordrecht, NL: Kluwer.
- Kohler, I. (1962). Experiments with goggles. *Scientific American*, 5 (206), 62-72.
- (1963). The formation and transformation of the perceptual world. *Psychological Issues*, 3 (4), 1-173.
- Kriegel, U. (2002). Phenomenal content. *Erkenntnis*, 57 (2), 175-198.
- (2007). Intentional inexistence and phenomenal intentionality. *Philosophical Perspectives*, 21 (1), 307-340. [10.1111/j.1520-8583.2007.00129.x](#)
- Levine, J. (1988). Absent and inverted qualia revisited. *Mind and Language*, 3 (4), 271-287. [10.1111/j.1468-0017.1988.tb00147.x](#)
- Lewis, C. (1929). *Mind and the world order*. New York, NY: Charles Scribner's Sons.
- Locke, J. (1689/1979). An essay concerning human understanding. In P. H. Nidditch (Ed.) *Clarendon edition*. Oxford, UK: Clarendon Press.

- Lupyan, G. (2015). Cognitive penetrability of perception in the age of prediction: Predictive systems are penetrable systems. *Review of Philosophy and Psychology*.
- Machery, E. (2009). *Doing without concepts*. Oxford, UK: Oxford University Press.
- Macpherson, F. (2005). Colour inversion problems for representationalism. *Philosophy and Phenomenological Research*, 70 (1), 127-152. [10.1111/j.1933-1592.2005.tb00508.x](https://doi.org/10.1111/j.1933-1592.2005.tb00508.x)
- (2012). Cognitive penetration of colour experience: Rethinking the issue in light of an indirect mechanism. *Philosophy and Phenomenological Research*, 84 (1), 24-62. [10.1111/j.1933-1592.2010.00481.x](https://doi.org/10.1111/j.1933-1592.2010.00481.x)
- Marks, L. E. & Odgaard, E. C. (2005). Developmental constraints on theories of synesthesia. In C. Robertson & N. Sagiv (Eds.) *Synesthesia: Perspectives from cognitive neuroscience* (pp. 214-236). New York, NY: Oxford University Press.
- McCauley, R. N. & Henrich, J. (2006). Susceptibility to the Müller-Lyer illusion, theory-neutral observation, and the diachronic penetrability of the visual input system. *Philosophical Psychology*, 19 (1), 79-101. [10.1080/095150805000462347](https://doi.org/10.1080/095150805000462347)
- McLaughlin, B. (2003). Color, consciousness, and color consciousness. *New essays on consciousness* (pp. 97-154). Oxford, UK: Oxford University Press.
- Meier, B. (2013). Semantic representation of synaesthesia. *Theoria et Historia Scientiarum*, 10, 125-134. [10.12775/ths-2013-0006](https://doi.org/10.12775/ths-2013-0006)
- Metzinger, T. (2003a). *Being no one: The self-model theory of subjectivity*. Cambridge, MA: MIT Press.
- (2003b). Phenomenal transparency and cognitive self-reference. *Phenomenology and the Cognitive Sciences*, 2 (4), 353-393. [10.1023/B:PHEN.0000007366.42918.eb](https://doi.org/10.1023/B:PHEN.0000007366.42918.eb)
- (2014). How does the brain encode epistemic reliability? Perceptual presence, phenomenal transparency, and counterfactual richness. *Cognitive Neuroscience*, 5 (2), 122-124. [10.1080/17588928.2014.905519](https://doi.org/10.1080/17588928.2014.905519)
- Montague, M. & Bayne, T. (2011). Cognitive phenomenology: An introduction. In M. Montague & T. Bayne (Eds.) *Cognitive phenomenology* (pp. 1-34). New York, NY: Oxford University Press.
- Mossio, M. & Taraborelli, D. (2008). Action-dependent perceptual invariants: From ecological to sensorimotor approaches. *Consciousness and Cognition*, 17 (4), 1324-1340. [10.1016/j.concog.2007.12.003](https://doi.org/10.1016/j.concog.2007.12.003)
- Mroczko, A., Metzinger, T., Singer, W. & Nikolić, D. (2009). Immediate transfer of synesthesia to a novel inducer. *Journal of Vision*, 9 (25), 1-8. [10.1167/9.12.25](https://doi.org/10.1167/9.12.25)
- Mroczko-Wąsowicz, A. (2013). *The unity of consciousness and phenomenon of synesthesia [Die Einheit des Bewusstseins und das Phänomen der Synästhesie]*. Published doctoral dissertation, Johannes Gutenberg University of Mainz.
- Mroczko-Wąsowicz, A. & Nikolić, D. (2013). Colored alphabets in bilingual synesthetes. *Oxford Handbook of Synesthesia* (pp. 165-180). Oxford, UK: Oxford University Press.
- (2014). Semantic mechanisms may be responsible for developing synesthesia. *Frontiers in Human Neuroscience*, 8 (509). [10.3389/fnhum.2014.00509](https://doi.org/10.3389/fnhum.2014.00509)
- Mroczko-Wąsowicz, A. & Werning, M. (2012). Synesthesia, sensory-motor contingency and semantic emulation: How swimming style-color synesthesia challenges the traditional view of synesthesia. *Frontiers in Psychology*, 3 (279). [10.3389/fpsyg.2012.00279](https://doi.org/10.3389/fpsyg.2012.00279)
- Myin, E. (2001). Color and the duplication assumption. *Synthese*, 129 (1), 61-77. [10.1023/A:1012647207838](https://doi.org/10.1023/A:1012647207838)
- Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83 (4), 435-450. [10.2307/2183914](https://doi.org/10.2307/2183914)
- Nanay, B. (2014). Cognitive penetration and the gallery of indiscernibles. *Frontiers in Psychology*, 5 (1527). [10.3389/fpsyg.2014.01527](https://doi.org/10.3389/fpsyg.2014.01527)
- Nelkin, N. (1989). Propositional attitudes and consciousness. *Philosophy and Phenomenological Research*, 49 (3), 413-430. [10.2307/2107796](https://doi.org/10.2307/2107796)
- Nida-Rümelin, M. (1993). *Farben und phänomenales Wissen. Eine Materialismuskritik. Conceptus-Studien 9*. Wien, AUT: Verlag der wissenschaftlichen Gesellschaften Österreichs.
- (1996). Pseudonormal vision. An actual case of qualia inversion? *Philosophical Studies*, 82 (2), 145-157.
- Nikolić, D., Lichti, P. & Singer, W. (2007). Color opponency in synaesthetic experiences. *Psychological Science*, 18 (6), 481-486. [10.1111/j.1467-9280.2007.01925.x](https://doi.org/10.1111/j.1467-9280.2007.01925.x)
- Noguchi, Y., Inui, K. & Kakigi, R. (2004). Temporal dynamics of neural adaptation effect in the human visual ventral stream. *The Journal of Neuroscience*, 24 (28), 6283-6290. [10.1523/JNEUROSCI.0655-04.2004](https://doi.org/10.1523/JNEUROSCI.0655-04.2004)
- Noë, A. (2005). *Action in perception*. Cambridge, MA: MIT Press.
- (2012). *Varieties of Presence*. Cambridge, MA: Harvard University Press.
- (2015). Concept pluralism, direct perception, and the fragility of presence. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- O'Regan, J. K. (2011). *Why red doesn't sound like a bell: Understanding the feel of consciousness*. Oxford, UK: Oxford University Press.

- O'Regan, J. K. & Noë, A. (2001). *A sensorimotor account of vision and visual consciousness*. Oxford, UK: Oxford University Press.
- Peacocke, C. (1983). *Sense and Content*. Oxford, UK: Oxford University Press.
- Pitt, D. (2004). The phenomenology of cognition or what is it like to think that P? *Philosophy and Phenomenological Research*, 69 (1), 1-36.
[10.1111/j.1933-1592.2004.tb.00382.x](https://doi.org/10.1111/j.1933-1592.2004.tb.00382.x)
- Prinz, J. J. (2010). Can concept empiricism forestall eliminativism? *Mind & Language*, 25 (5), 5612-5621.
[10.1111/j.1468-0017.2010.01404.x](https://doi.org/10.1111/j.1468-0017.2010.01404.x)
- (2012). *The conscious brain*. Oxford, UK: Oxford University Press.
- Proulx, M. J. & Störig, P. (2006). Seeing sounds and tingling tongues: Qualia in synaesthesia and sensory substitution. *Anthropology & Philosophy*, 7, 135-150.
- Pylyshyn, Z. W. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences*, 22 (3), 341-365. [10.1017/s0140525x99002022](https://doi.org/10.1017/s0140525x99002022)
- Raftopoulos, A. (2005). Cognitive penetrability of perception: A new perspective. *Cognitive penetrability of perception: Attention, action, strategies, and bottom-up constraints* (pp. 1-25). Hauppauge, NY: Nova Science.
- (2009). *Cognition and perception: How do psychology and the neural science inform philosophy*. Cambridge, MA: MIT Press.
- Ramachandran, V. S. & Hubbard, E. M. (2001a). Synaesthesia: A window into perception, thought and language. *Journal of Consciousness Studies*, 8, 3-34.
- (2001b). Psychophysical investigations into the neural basis of synaesthesia. *Proceedings of the Royal Society of London B: Biological Sciences*, 268 (1470), 979-983. [10.1098/rspb.2000.1576](https://doi.org/10.1098/rspb.2000.1576)
- Ramachandran, V. S. & Rogers-Ramachandran, D. (1996). Synaesthesia in phantom limbs induced with mirrors. *Proceedings of the Royal Society of London B: Biological Sciences*, 263 (1369), 377-386.
[10.1098/rspb.1996.0058](https://doi.org/10.1098/rspb.1996.0058)
- Rees, G. & Frith, C. (2007). Methodologies for identifying the neural correlates of consciousness. In M. Velmans & S. Schneider (Eds.) *The Blackwell companion to consciousness* (pp. 553-566). Malden, MA: Blackwell.
- Robinson, H. (1994). *Perception*. London, UK: Routledge.
- Schwitzgebel, E. (2008). The unreliability of naïve introspection. *The Philosophical Review*, 117 (2), 245-273.
[10.1215/00318108-2007-037](https://doi.org/10.1215/00318108-2007-037)
- Seitz, A. & Watanabe, T. (2005). A unified model for perceptual learning. *Trends in Cognitive Sciences*, 9 (7), 329-334. [10.1016/j.tics.2005.05.010](https://doi.org/10.1016/j.tics.2005.05.010)
- Seth, A. K. (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synaesthesia. *Cognitive Neuroscience*, 5 (2), 97-118.
[10.1080/17588928.2013.877880](https://doi.org/10.1080/17588928.2013.877880)
- (2015). The cybernetic Bayesian brain: From interoceptive inference to sensorimotor contingencies. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Shoemaker, S. (1982). The inverted spectrum. *Journal of Philosophy*, 79 (7), 357-381.
- (1994). Phenomenal character. *Noûs*, 28 (1), 21-38. [10.2307/2215918](https://doi.org/10.2307/2215918)
- (2001). Introspection and phenomenal character. *Philosophical Topics*, 28 (2), 247-273.
[10.5840/philtopics20002825](https://doi.org/10.5840/philtopics20002825)
- Siegel, S. (2006). Which properties are represented in perception? In T. Gendler & J. Hawthorne (Eds.) *Perceptual Experience* (pp. 481-503). Oxford, UK: Oxford University Press.
- (2012). Cognitive penetrability and perceptual justification. *Noûs*, 46 (2), 201-222.
[10.1111/j.1468-0068.2010.00786.x](https://doi.org/10.1111/j.1468-0068.2010.00786.x)
- Siewert, C. (1998). *The significance of consciousness*. Princeton, NJ: Princeton University Press.
- Simner, J. (2007). Beyond perception: Synaesthesia as a psycholinguistic phenomenon. *Trends in Cognitive Sciences*, 11 (1), 23-29. [10.1016/j.tics.2006.10.010](https://doi.org/10.1016/j.tics.2006.10.010)
- (2012). Defining synaesthesia: A response to two excellent commentaries. *British Journal of Psychology*, 103 (1), 24-27. [10.1111/j.2044-8295.2011.02059.x](https://doi.org/10.1111/j.2044-8295.2011.02059.x)
- Smithson, H. E. (2005). Review. Sensory, computational and cognitive components of human color constancy. *Philosophical Transactions of the Royal Society*, 360 (1458), 1329-1346.
[10.1098/rstb.2005.1633](https://doi.org/10.1098/rstb.2005.1633)
- Stoerig, P. & Cowey, A. (1989). Wavelength sensitivity in blindsight. *Nature*, 342 (6252), 916-918.
[10.1038/342916a0](https://doi.org/10.1038/342916a0)
- (1991). Increment threshold spectral sensitivity in blindsight: Evidence for colour opponency. *Brain*, 114 (3), 1487-1512.
- Stokes, D. (2012). Perceiving and desiring: A new look at the cognitive penetrability of experience. *Philosophical Studies*, 158 (3), 479-492.
[10.1007/s11098-010-9688-8](https://doi.org/10.1007/s11098-010-9688-8)

- Strawson, P. F. (1985). Causation and explanation. In B. Vermazen & J. Hintikka (Eds.) *Essays on Davidson* (pp. 115-135). Oxford, UK: Oxford University Press.
- (1994). *Mental reality*. Cambridge, MA: MIT Press.
- Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18 (6), 643-662. [10.1037/h0054651](https://doi.org/10.1037/h0054651)
- Taraborelli, D. & Mossio, M. (2008). On the relation between the enactive and the sensorimotor approach to perception. *Consciousness and Cognition*, 17 (4), 1343-1344. [10.1016/j.concog.2008.08.002](https://doi.org/10.1016/j.concog.2008.08.002)
- Tye, M. (1993). Qualia, content, and the inverted spectrum. *Noûs*, 27 (2), 159-183. [10.2307/2216047](https://doi.org/10.2307/2216047)
- (1995). Ten problems of consciousness. *Consciousness, color and content*. Cambridge, MA: MIT Press.
- (2000). *Consciousness, Color and Content*. Cambridge, MA: MIT Press.
- van Leeuwen, T. M., Petersson, K. M. & Hagoort, P. (2010). Synaesthetic colour in the brain: Beyond colour areas. A functional magnetic resonance imaging study of synaesthetes and matched controls. *PLoS ONE*, 5, e12074.
- Varela, F. J. (1996). Neurophenomenology: A methodological remedy for the hard problem. *Journal of Consciousness Studies*, 3 (4), 330-349.
- Vetter, P. & Newen, A. (2014). Varieties of cognitive penetration in visual perception. *Consciousness and Cognition*, 27, 62-75. [10.1016/j.concog.2014.04.007](https://doi.org/10.1016/j.concog.2014.04.007)
- Vuilleumier, P. & Driver, J. (2007). Modulation of visual processing by attention and emotion: Windows on causal interactions between human brain regions. *Philosophical transactions of the Royal Society of London. Series B: Biological sciences*, 362 (1481), 837-855. [10.1098/rstb.2007.2092](https://doi.org/10.1098/rstb.2007.2092)
- Ward, D. (2012). Why don't synaesthetic colours adapt away? *Philosophical Studies*, 159 (1), 123-138. [10.1007/s11098-010-9693-y](https://doi.org/10.1007/s11098-010-9693-y)
- Webster, M. A. (2012). Evolving concepts of sensory adaptation. *F1000 Biology Reports*, 4 (21). [10.3410/B4-21](https://doi.org/10.3410/B4-21)
- Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R. & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences USA*, 104 (19), 7780-7785. [10.1073/pnas.0701644104](https://doi.org/10.1073/pnas.0701644104)
- Wittgenstein, L. (1953). *Philosophical investigations*. Oxford, UK: Basil Blackwell.
- Wu, W. (2013). Visual spatial constancy and modularity: Does intention penetrate vision? *Philosophical Studies*, 165, 647-669. [10.1007/s11098-012-9971-y](https://doi.org/10.1007/s11098-012-9971-y)
- Zmigrod, S. & Hommel, B. (2011). The relationship between feature binding and consciousness: Evidence from asynchronous multi-modal stimuli. *Consciousness and Cognition*, 20 (3), 586-593. [10.1016/j.concog.2011.01.011](https://doi.org/10.1016/j.concog.2011.01.011)

Phenomenology, Methodology, and Advancing the Debate

A Reply to Aleksandra Mroczko-Wąsowicz

Rick Grush

The following topics are briefly discussed: First, the different senses of what counts as phenomenal, and in particular how this might influence how our results are described; second, the methodological limitations of our original study; and finally, some ways that the commentary by Mroczko-Wąsowicz charts out potential theoretical advancement of the results we presented in our study.

Keywords

Methodology | Phenomenology | Subjective report | Synesthesia

Author

[Rick Grush](#)

rick@mind.ucsd.edu

University of San Diego
San Diego, CA, U.S.A.

Commentator

[Aleksandra Mroczko-Wąsowicz](#)

mroczko-wasowicz@hotmail.com

國立台灣大學

National Yang Ming University
Taipei, Taiwan

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 On the nature of phenomenology

First, as Mroczko-Wąsowicz quite rightly points out, there are different understandings of what *phenomenology* is, with concomitant differences in what *phenomenal adaptation* might mean. The distinction drawn is between phenomenal conservatism, and phenomenal liberalism; the former being constrained to the vicinity of sensory features, and the latter including various cognitive phenomena, such as expectations and associations, among others.

We chose to use the term in the more restrictive sense for a number of reasons. First, as the more restrictive of the two, it is less controversial that what is included counts as genuinely phenomenological. Second, in many circles at least, the more restrictive understanding seems to be what people generally have in mind. The more liberal understanding is one that is endorsed more commonly only among specialists.

What I am about to say may be a matter of splitting hairs – and so I ask for forgiveness

in advance. I am in complete agreement that the distinction is a valuable one to make, and that in our original article we just ran with the more restrictive definition. That said, it doesn't seem to me that with this distinction in hand one is raising "alternative interpretations" of our results; rather one is providing a different way of *describing* the same result. On a conservative definition of what counts as phenomenal, we did not find phenomenal adaptation. But if one adopts a liberal understanding of the term that includes various cognitive phenomena, then it would be correct to say that we did, in fact, find some phenomenal adaptation. So long as there is clarity on what exactly was found, and on how one intends to use the key terms, then this shouldn't be cause for confusion or concern.

Where things could get interesting would be on a possible third way to understand phenomenal – call it the *radical* understanding. On the radical view, there is nothing to phenomenology other than the sort of cognitive phenomena that the liberal view intended to add to the more narrowly sensory understanding. For one who holds such a view, we may very well have found the beginnings of phenomenal adaptation *tout court* when we found the beginnings of elements of cognitive adaptation.

This hairsplitting aside, I couldn't agree more with Mroczko-Wąsowicz's point that when getting into the details of discussions about phenomenal adaptation, a solid understanding of the different ways that the key terms might be understood is crucially important, and in this respect her commentary is an excellent supplement to the discussion we provided.

2 On methodological limitations

Mroczko-Wąsowicz goes on to, quite reasonably, point to some of the shortcomings of our pilot study. In fact, we pointed out many of these same shortcomings ourselves. There are a couple however that are worth saying at least a bit about.

Mroczko-Wąsowicz points out that some of our findings are based on subjective report, and that there are "doubts about whether subjective reports are trustworthy." While in general this is

entirely correct, there is a sense of phenomenal adaptation according to which what we were studying is precisely *how things would seem to the subject*. It is undoubtedly the case that even in such situations one is not *limited to* what subjective report might have to say on the matter. Indeed, this is among the reasons we included other experiments as part of the protocol. But the phenomenon that I *subjectively notice and can report on* when I adapt to the spatial distortion of new corrective glasses, or to the color distortion of blue-blocking sunglasses is an interesting one, and one might reasonably wonder if one can get an analogous adaptation effect – the same subjectively noticeable and reportable effect – with respect to rotated colors.

This is related to a second point. Mroczko-Wąsowicz echoes our claim that it is a shortcoming of the study that the researchers themselves were subjects. Surely it is the case that knowledge of the experiment and the phenomena to be studied can bias the results. Of course I agree completely with that.

Nevertheless, I am reminded of a point made in conversation by Vilayanur Ramachandran. In a moment of venting about some objections made to some of his results, he hypothesized that he could show psychologists a talking pig and they would scoff that it was an n of 1.

In the present case, it is true that having the experimenters themselves be subjects effects the results. But even so, if it turned out that I or the other subject JK did end up in a state that seemed to us to be one of phenomenal adaptation, then this would still be interesting, because if nothing else it would demonstrate that we could get the effect in anyone if we just briefed them on the experiment beforehand. If I hypothesize that hitting myself on the head three times with a baguette will make me able to speak fluent French, and I do the experiment and it *does*, this is an interesting result even if I was both experimenter and subject.

In any case, Mroczko-Wąsowicz and I are in a great deal of agreement about the limitations created by the methodology of our pilot study, and these limitations need to be kept firmly in mind when anyone ventures to interpret our findings or follow up on them.

3 Advancing the debate

The final set of points made by Mroczko-Wąsowicz concerns synesthesia, and in particular how the phenomenon presents an interesting complement to the sort of phenomenon we studied. When we were initially brainstorming the experiment we discussed what might happen if a synesthete were to wear the rotation gear. But that line of speculation never got past the brainstorming stage, since just doing it with ourselves proved enough of a challenge. While it has some significant differences from synesthesia, we did make an attempt to see whether the McCollough effect would adapt. But the subjective effect was very small, and didn't last long enough into the protocol to get any data at the time when there might have been some adaptation.

Mroczko-Wąsowicz makes some fascinating points about how our study and synesthesia complement each other in interesting way that would be strong motivation for anyone following up on our study to try to include some synesthetes among the test subjects.

4 Conclusion

We tried to make our initial article streamlined and not burdened with too much detailed theoretical discussion. Since we hope the interested parties will include not only philosophers but also psychologists and cognitive scientists, the thought was to present the results, which we thought were quite interesting and suggestive, and leave the more detailed theoretical discussions and possible follow-up experiments to others. In this respect Mroczko-Wąsowicz's commentary is exactly the sort of detailed theoretical follow-up we hoped others might be inspired to produce on the basis of our results. I am grateful to her for fantastic commentary.

An Information-Based Approach to Consciousness: Mental State Decoding

John-Dylan Haynes

The debate on the neural correlates of visual consciousness often focuses on the question of which additional processing has to happen for a visual representation to enter consciousness. However, a related question that has only rarely been addressed is which brain regions directly encode specific contents of consciousness. The search for these core neural correlates of contents of consciousness (NCCCs) requires establishing a mapping between sensory experiences and population measures of brain activity in specific brain regions. One approach for establishing this mapping is multivariate decoding. Using this technique, several properties of NCCCs have been investigated. Masking studies have revealed that information about sensory stimuli can be decoded from the primary visual cortex, even if the stimuli cannot be consciously identified by a subject. This suggests that information that does not reach awareness can be encapsulated in early visual stages of processing. Visual imagery representations and veridical perception share similar neural representations in higher-level visual regions, suggesting that these regions are directly related to the encoding of conscious visual experience. But population signals in these higher-level visual regions cannot be the sole carriers of visual experiences because they are invariant to low-level visual features. We found no evidence for increased encoding of sensory information in the prefrontal cortex when a stimulus reaches awareness. In general, we found no role of the prefrontal cortex in encoding sensory experiences at all. However, the improved discrimination of sensory information during perceptual learning could be explained by an improved read-out by the prefrontal cortex. One possible implication is that prefrontal cortical regions do not participate in the encoding of sensory features per se. Instead they may be relevant in making decisions about sensory features, without exhibiting a re-representation of sensory information.

Keywords

Imagery | Masking | Multivariate decoding | Neural correlate of consciousness | Sensory information

1 Introduction

Neural theories of visual consciousness frequently focus on the question of what is needed for a visual stimulus to enter consciousness. A common notion is that representations and processes in sensory regions of the brain can operate outside of conscious perception, and that some “extra property of processing” has to come on top in order to let these representations enter conscious experience (e.g., [Dehaene & Naccache 2001](#)). This extra processing property can range from neural synchronization of neurons encoding the stimulus ([Engel & Singer 2001](#)), recurrent and feedback processing ([Lamme 2006](#), [this collection](#); [Pascual-](#)

[Leone & Walsh 2001](#); [Singer this collection](#)), to participation in a global coherent process, known as neuronal workspace theories ([Baars 2002](#); [Dehaene & Naccache 2001](#)). Discussion of the neural correlates of consciousness (NCC) has often focused on this extra ingredient needed to bring a stimulus representation into consciousness. However, a related, but somewhat different question has often been neglected: Which neural representations (can) precisely participate in encoding the various dimensions of conscious experience? For this it is not enough to establish a correlation between conscious perception and neural signals.

Author

[John-Dylan Haynes](#)

haynes@bccn-berlin.de

Charité – Universitätsmedizin
Berlin, Germany

Commentator

[Caspar M. Schwiedrzik](#)

cschwiedrz@rockefeller.edu

The Rockefeller University
New York, NY, U.S.A.

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

Glossary

Neural encoding	The representation of a sensory feature in a population of neurons.
Mental state decoding	Inferring the representational content of a mental state from a brain activation pattern, typically using multivariate pattern classification.
Neural correlate of a content of consciousness	The brain signal that encodes a specific aspect of conscious experience. The brain signal is the carrier, the specific aspect of consciousness constitutes the phenomenal representational content carried (in short, its “phenomenal content”).
Multivariate pattern classification	A mathematical procedure for identifying patterns of brain activity, the labels of which have been previously learned.
Mapping	The assignment of brain activation patterns to the representational content of mental states.
Low-level visual features	Simple dimensions of visual experience that are encoded in early visual brain regions (e.g., contrast, orientation). If consciously represented, they may constitute corresponding simple forms of phenomenal content.
High-level visual features	More complex dimensions of visual experience that are encoded in downstream visual brain regions (e.g., object identity) and that are to some degree independent of the low-level features by which they are defined.

That would yield a far too large set of candidate brain regions, including, say, signal patterns in the retina that also correlate with conscious perception. Instead, it would be desirable to identify which neural representations most closely encode specific contents of consciousness and can be used to explain dimensions of conscious perception under as many different conditions as possible, and down to the level of single contents. This article will focus on how to identify such core neural correlates of contents of consciousness (NCCCs; [Chalmers 2000](#); [Block 2007](#); [Koch 2004](#)).

It is desirable that studies of visual awareness take NCCCs into account because specific theories of visual awareness make specific predictions regarding the encoding and distribution of sensory information (e.g., [Dehaene & Naccache 2001](#); see also [Baars 2002](#)). In the following, I will first outline the more standard techniques for identifying NCCCs, along with their shortcomings. The next step proposes to use multivariate decoding techniques (reviewed e.g., in [Haynes & Rees 2006](#)) as a tool to identify NCCCs. Decoding can serve as an empirical technique that can establish which brain regions bear most information about specific contents of visual experience. This is an important first step towards establish-

ing a more rigid mapping between visual phenomenal states and content-encoding brain signals. Then, several examples will be presented where multivariate decoding of visual experiences can help inform specific questions regarding NCCCs, such as whether information in V1 participates in visual awareness, whether imagery and perception share the same underlying neural codes, or whether the prefrontal cortex contains any dynamic NCCCs for coding specific dimensions of conscious experience.

2 Why content matters: Binocular rivalry and the multiple levels of conscious experience

In 1996 [Nikos Logothetis & David Leopold](#) published a landmark study on the neural mechanisms of visual awareness. They presented their participating monkeys with a very elaborate visual stimulus display, which allowed them to show one image to one eye and another image to the other eye. For example, the left eye might be stimulated with a line pattern tilted to the left, and the right eye might be stimulated with a line pattern tilted to the right. In such cases, where conflicting input is presented to



Figure 1: Binocular rivalry and levels of perception. (a) Two conflicting stimuli, one presented to the left and one to the right eye, lead to a perceptual alternation between phases where the input of either the left or right eye is consciously seen. In monkey single-cell electrophysiology, this perceptual alternation has a correlate in higher-level visual regions of the temporal lobe, but activity in earlier visual regions shows only small changes in activity patterns. Presumably, signals in the temporal cortex encode the complex figural properties of the stimuli, such as the left being a sunburst pattern and the right being an image of a monkey face. However, due to the invariance of brain responses in higher-level visual regions to low-level features, this cannot explain the perceptual difference between the rivalry of the left and of the right sunburst pattern shown in (b), where the central circle has changed colour but the entire shape remains similar (monkey illustration by Chris Huh, Wikimedia Commons).

the two eyes, human participants don't experience a fusion between the two images. Instead, conscious visual perception alternates between phases where one of the eyes' inputs become visible and phases where the other eye's input are seen. Perception waxes and wanes more or less randomly between two perceptual experiences—despite constant stimulation. Similarly, the monkeys that were exposed to these binocular rivalry stimuli indicated behaviorally by pressing levers that their perception alternated between the inputs to the two eyes.

In parallel, Leopold & Logothetis (1996) investigated what happened to the firing patterns of single neurons in the monkeys' brains. Their setup allowed them to not just look at one location in the brain, but to assess neural correlates in several visual brain regions. They found only a small percentage of single neurons in early visual cortex (V1/V2) whose firing pattern was modulated by the stimulus that was currently dominant. In contrast, in a higher-level visual area—V4—they found that many more cells changed their firing rates with changes in perception. This establishes a clear dissociation between early visual areas where neural signals seem not to correlate with awareness and high visual areas where they do. In a follow-up experiment, Sheinberg & Logothetis

(1997) investigated the involvement of even higher visual regions in the temporal cortex in binocular rivalry. Because cells in these regions preferentially respond to more elaborate visual features, they used complex shapes and images, such as, for example, an abstract sunburst pattern or a picture of a monkey face (Figure 1a). They found that in the superior temporal sulcus and in the inferior temporal cortex, a large percentage of cells modulated their firing rate with perceptual dominance. Taken together, these studies seem to suggest that visual awareness affects signals only at late stages of the visual system.

But what does it mean exactly that visual awareness only affects late stages of visual processing? Does it mean that high level visual areas contain all the neural correlates of contents consciousness (NCCCs), in a way similar to a CD encoding the contents of a piece of music? If the signals in these high visual areas are really responsible for encoding all contents of visual experiences then any aspect of conscious perception that changes during binocular rivalry should be explainable by changes in signals in these higher-level brain regions. There are reasons to believe that this cannot be the case. Consider the two images as shown in Figure 1a. At one instant the monkey might

consciously see the face image. This percept would be encoded in activity patterns in the higher visual areas. In the next instant the monkey might see a sunburst pattern, and this experience would also be encoded in the higher visual cortex. At first sight this seems reasonable. Higher-level visual areas are specialized for complex visual information and object features (Sáry et al. 1993). So cells that have a preference for faces might respond during dominance of the face image, and cells with a preference for sunburst patterns might respond during the dominance of that pattern. But there is one difficulty in this interpretation. The images have a high-level interpretation as complex shapes, but they are also composed of a multitude of minute visual features, edges, surfaces, colours, etc. During rivalry, our perception does not only change according to the abstract interpretation, with respect to abstract, high-level interpretation, but also in terms of the minute, fine-grained details of visual experience (see Figure 1b).

This poses a problem because responses in higher-level visual areas are invariant with respect to low-level features (Sáry et al. 1993). Cells in higher-level visual areas in the inferior temporal cortex respond selectively to specific object features in an invariant pattern (Figure 2). A cell specialized for detecting, say, a circle, will respond to this circle irrespective of the low-level features by which it is defined (here brightness, contrast, and colour contrast). This means that such a cell disregards the low-level features and does not convey information about them any more. While cells in high-level visual areas might be able to explain why we see a face one moment and a sunburst pattern the next, they cannot explain why the sunburst pattern is yellow instead of red, or why it is one specific visual pixel collection out of the many possible that would be seen as a sunburst pattern. Thus, visual experience is a multilevel phenomenon, and a theory of the neural correlates of visual awareness will have to be able to explain all the levels of our experience, not just one. This clearly shows the importance of a content-based approach to visual consciousness.

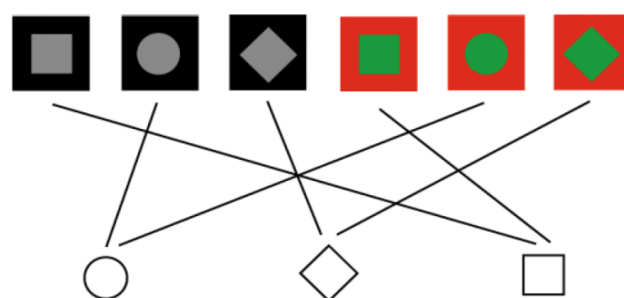


Figure 2: Invariance of single-cell responses in higher-level visual areas. Responses in low-level visual areas (top) are tuned to low-level features such as colour or luminance. In contrast, responses in higher-level object-selective regions (bottom) are largely invariant with respect to these low-level features (Sáry et al. 1993).

3 Approaches to content-selectivity in human neuroimaging and their problems

The limited resolution of current neuroimaging techniques poses a substantial problem for the investigation of the encoding of contents in the human brain, and thus for studies on NCCCs in the human brain. The most important format in which information is coded in the brain is the cortical column (Fujita et al. 1992). Cortical columns consist of small groupings of cells with similar tuning properties, clustered together at a scale of around half a millimetre. Even functional magnetic resonance imaging (fMRI) does not routinely have a sufficient resolution to selectively study the activation of individual cortical columns (but see e.g., Yacoub et al. 2008 for recent progress). For this reason, most research into perceptual contents has relied on experimental “tricks” that allow the tracking of contents indirectly.

In *frequency tagging*, a visual stimulus is tagged with a specific and unique flicker frequency. This then allows for tracing of the processing of this stimulus by searching for brain signals that exhibit the same flicker frequency. This approach has been used to study binocular rivalry, but in quite a different way to that undertaken by Leopold & Logothetis (1996). Tononi et al. (1998) tagged the inputs of the two eyes with different frequencies. They found that the currently dominant percept was accom-

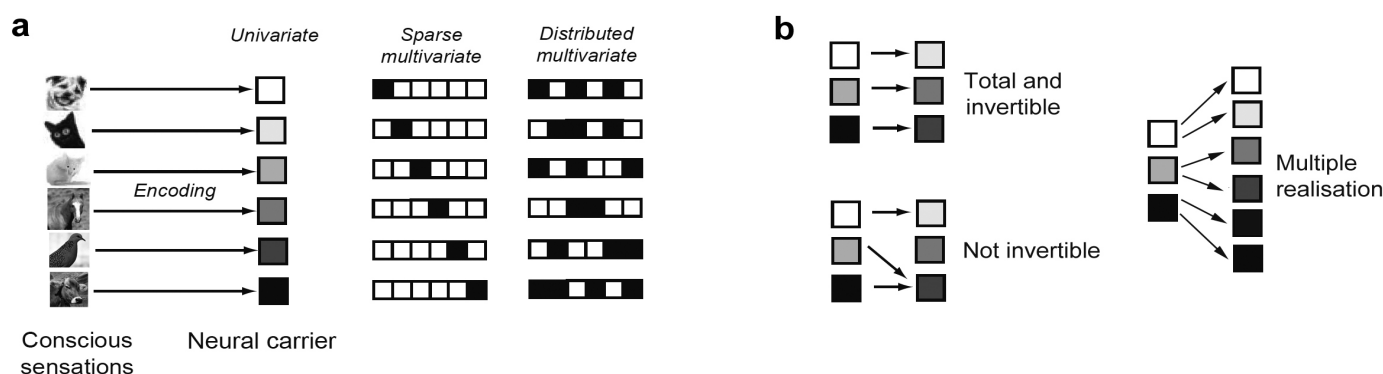


Figure 3: Principles of mapping between mental states and brain states (see text; adapted from Haynes 2009 with additional images from Wikipedia).

panied by wide-spread increases in Magnetoencephalography (MEG)-signals at the tagged frequency across multiple brain regions, mostly in the early visual and temporal cortex. This is a very powerful approach and it reveals how wide-spread the effects are when a stimulus reaches visual awareness. However, it is not always clear whether these findings indicate that the corresponding perceptual features of the stimuli, in this case the orientation of line elements, really are distributed throughout the brain. The key problem is that the feature that is traced (the frequency) is not the main feature that is perceptually relevant (orientation). One could imagine, say, that activity in higher-level brain regions that exhibits the frequency of the dominant stimulus might not be involved in coding the sensory content, but instead in detecting the presence of a change in the visual image, irrespective of what the corresponding feature is. The frequency-tagging approach does not allow for distinguishing between these alternatives.

Another approach to tracking content-selective processing is to use stimuli that are known to activate specific *content-selective brain regions* (Tong et al. 1998; Rees et al. 2000). For example, in a study on binocular rivalry, Tong et al. (1998) used faces and houses as rivalry stimuli. These stimuli are known to activate different brain regions, the fusiform face area (FFA) and the parahippocampal place area (PPA). They found that activity in a content-selective region increased when the corresponding stimulus became perceptually dominant. This goes further than the frequency-tag-

ging approach in that it allows for drawing the plausible conclusion that awareness leads to increased activity in content-selective regions. However, this approach again suffers from several problems. First, it only allows us to address the hypothesis related to very specific stimuli (typically faces and houses) and to very specific brain regions. Because the approach relies on the existence of macroscopic content-selective regions, it would not be possible to test whether, say, the prefrontal cortex, receives sensory information when a stimulus reaches awareness. A further problem is that the high selectivity of FFA and PPA has long been questioned (Haxby et al. 2001).

4 Mapping and decoding

It seems a different, more direct, and generic approach is necessary in order to identify the neural correlates of the contents of consciousness. It may help to start in the simplest possible way. If we want to explain the occurrence of a conscious experience E_1 with the occurrence of a brain state B_1 then—roughly speaking—the experience and the brain state should always happen together. If we want to explain N experiences $E_{1..N}$ with brain states, we will need N different brain states $B_{1..N}$ in order to encode the different experiences. If brain data from a specific area only adopts one of five states every time a participant has one of ten experiences it is impossible to explain the experiences through the different brain states. Ultimately this boils down to a *mapping* problem (Haynes 2009; Figure 3).

A set of conscious sensations, here visual percepts of six different animals, can be encoded in a neural carrier in multiple ways. Three principles are illustrated in [Figure 3a](#). One hypothetical way to code these six animals would be to use a single neuron and to encode the objects by the firing rates of this neuron. One would assign one specific firing rate to each of the different animals, say 1Hz to the dog, 2Hz to the cat and 3Hz to the mouse, etc. This approach is also referred to as a *univariate* code, because it uses only one single parameter of neural activity. It has the advantage of requiring only a single neuron. In *principle* it is possible to encode many different objects with a single neuron. The idea would be very similar to a telephone number, if one thinks of different numbers corresponding to different firing rates. In theory it would be possible to encode every single telephone in the world in this way, provided that the firing rates could be established very precisely and reliably. The disadvantage with this approach—even if firing rates could be established with great precision—is that it can only handle *exclusive* thoughts, i.e. it has no way of dealing with a superposition of different animals, say a cat together with a dog.

A different approach is not to use a single neuron to encode different thoughts, but instead to use a set of neurons to encode a set of thoughts. This population-based approach is also termed “multivariate”. One way to encode thoughts about six different animals would be to assign one specific neuron to the occurrence of each thought. Neuron one, say, might fire when a person thinks about a dog; neuron two would fire when they were thinking about a cat, etc. Here the firing rate is irrelevant; only a threshold is needed, such that one has a way of deciding when a neuron is “active” or “not active”. This specific coding scheme is variably termed “sparse code”, “labelled line code”, “cardinal cell code” or “grandmother cell code” (see e.g., [Quiroga et al. 2008](#)). It has the advantage of being able to handle arbitrary superpositions and combinations of thoughts, say thoughts about a meeting of a dog, a cat, and a mouse. A disadvantage is that a different neuron is needed for the encoding of each new entity. N

neurons can only encode N different thoughts. Given that the average human brain comprises 86 billion neurons ([Azevedo et al. 2009](#)) this might not seem too big a problem. A different way to use a population of neurons to encode a set of thoughts would be a *distributed* multivariate code. Here, each mental state is associated with a single activation pattern in the neural population, but now arbitrary combinations of neurons are possible for the encoding of each single thought. This allows for the encoding of 2^N thoughts with N neurons, if each neuron is only considered to be “on” or “off”.

There are various examples of these different types of codes. The encoding of *intensity* follows a univariate code: The difference between a brighter and a darker image is encoded in a higher versus a lower firing rate of the corresponding neurons in the visual cortex (see e.g., [Haynes 2009](#)). However, to date, I am not aware of any example where different higher-level interpretations of stimuli are coded in a univariate format. There are many examples of labelled line codes. The retinotopic location within the visual field is encoded in a sparse, labelled line format (e.g., [Serenio et al. 1995](#)). One position in the visual field is coded by one set of neurons in the early visual cortex; another position is encoded by a different set of neurons. If two objects appear in the visual field simultaneously, then both of the corresponding sets of neurons become active simultaneously. A similar coding principle is observed for auditory pitch, where different pitches are coded in different cells in the form of a tonotopic map ([Formisano et al. 2003](#)). The somatosensory and motor homunculi are also examples of labelled line codes, each position in the brain corresponding to one specific position in the body ([Penfield & Rasmussen 1950](#)). A distributed multivariate code is, for example, used to code different objects ([Haxby et al. 2001](#)) or different emotions ([Anders et al. 2011](#)).

When identifying the mapping between brain states and mental states one is generally interested in identifying which specific population of neurons is a suitable candidate for explaining a particular class of visual experiences. For this it is possible to formulate a number of

constraints (Haynes 2009). First, the mapping needs to assign one brain state to each mental state in which we are interested. In other words, the mapping has to be total (Figure 3b). This should be easy—it just means that we can assign one measured brain state to each different mental state. Second, the mapping cannot assign the same brain state to two different mental states. Otherwise the brain states would not be able to explain the different mental states. Technically this means the mapping has to be invertible, or injective. Every brain state should be assigned to no more than one mental state. However, it is possible—in the sense of multiple realisation—that multiple brain states are assigned to the same mental state, as long as neither of these brain states co-occurs with other mental states. The brain states referred to here only mean brain states that are relevant for explaining a set of mental states. If we want to explain thoughts about six animals, say, it might not be necessary that brain states in the motor cortex are different for the different animals. However, if one wants to propose one set of neurons (say, those in the lateral occipital complex, Malach et al. 1995) as a candidate for explaining animal experiences, then this can only hold if the abovementioned mapping requirements are fulfilled.

In practice it will be very difficult to establish this mapping directly. One major problem is that we can't measure brain states in sufficient detail with current neuroscience techniques. Non-invasive measurement techniques such as electroencephalography (EEG) or functional magnetic resonance imaging (fMRI) have very limited spatial resolution. fMRI for example resolves the brain with a measurement grid of around 1–3mm, so that each measurement unit (or voxel) contains up to a million cells. And the temporal resolution of fMRI is restricted because fMRI measures the delayed and temporally-extended hemodynamic response to neural stimulation. While it is possible—to some degree—to reconstruct visual experiences from fMRI signals (e.g., Miyawaki et al. 2008), fMRI cannot resolve temporal details of neural processes, such as the synchronized activity of multiple cells. But it is not only EEG and fMRI

that have limited resolution: Invasive recording techniques are typically restricted to individual well-circumscribed locations, where surgery is performed. And even with multielectrodes it is not possible to identify the state of each individual neuron in a piece of living tissue.

Another important limitation lies in our ability to precisely characterize and cognitively penetrate *phenomenal states* (e.g., Raffman 1995). There is currently no psychophysical technique that would allow us to characterize the full details of a person's visual experiences at each location in the visual field. Verbal reports or button presses can convey only a very reduced picture of the true complexity of visual experiences. So ultimately, the mapping requires precision from both psychology and neuroscience, and any imprecision in either approach will blur the mapping and distort the interpretation.

The next best option short of establishing full mapping is to use decoding techniques that follow a similar logic. Brain-based decoding is also referred to as “brain reading” or “multivoxel pattern analysis” (see Haynes & Rees 2005 for a review). The basic idea is to see to which degree it is possible to infer a mental state from a measurement of a brain state. Say you want to test whether the lateral occipital complex is a suitable candidate for encoding visual thoughts about animals. You test if it is possible to infer which animal a person is currently seeing by training a classifier to learn the association between animal and brain activation pattern, and then one needs to test whether the classifier can correctly assign the animal that belongs to a new measurement of brain activity. In the following, this approach will be explained in detail.

Take for example a hypothetical fMRI-measurement of the human brain within a three by three grid of locations, amounting to nine voxels (Figure 4a). These nine voxels can be systematically resorted into a column of numbers (or vectors), where each entry denotes the activation at one location (high values correspond to strong fMRI responses). Say one was interested in testing whether these nine voxels contain information about two different visual

images, perhaps a dog and a cat. The question that needs to be addressed is whether the response patterns (i.e., the vectors) are sufficiently different to allow for distinguishing between the animals, based on these brain activity measurements alone. The vector is not a useful way to see whether this classification is possible. It can help to visualize the same information in a two-dimensional coordinate system. Take the responses to the dog. One can think of the first and second entries in the vector as x- and y-values that define points in a coordinate system. The response in the first voxel (x) to the dog is a low value (2), while the second value (y) is a high value (8). When plotted in a two-dimensional coordinate system (Figure 4b), this yields a point in the top left of the coordinate system, shown here in red. Repeated measurements of the brain response to the dog yield a small cloud of red points. Repeatedly measured brain responses to the cat have high values in voxel 1 (x) and low values in voxel 2 (y). In the two-dimensional coordinate system this yields a cloud of blue points in the bottom right of the coordinate system. Clearly the responses are separable in this two-dimensional coordinate system, so the two animals enjoy reliably separate neural representations in this set of nine voxels. In this hypothetical example, each of the two voxels alone would be sufficiently informative about the category of animal. By collapsing the points for voxel one onto the x-axis it becomes clear that the two distributions of points (red and blue) are sufficiently different to allow for telling the two apart. The same holds for voxel two by collapsing to the y-axis. This is akin to a labelled line code, with one line for “dog” and one line for “cat”.

However, there are cases where the two distributions will not be so easily separable. Figure 4c shows an example where the individual voxels do not have information about the animals. The collapsed or “marginal” distributions largely overlap. There is no way to tell a cat response from a dog response by looking at either voxel one or two alone. However, by taking into account the *joint activity* in both voxels, the two animals become clearly separ-

able. Responses to the dog all cluster to the top left of the diagonal and responses to the cat cluster to its bottom right. This joint consideration of the information contained in multiple voxels is the underlying principle of *multivariate decoding*. The line separating the two distributions is known as the decision boundary. Decision boundaries are not necessarily straight lines. Many other types of distributions of responses are possible. Figure 4d, for example, shows a non-linear decision boundary. Finding the optimal decision boundary is the key objective in the field of machine learning (Müller et al. 2001), where many different types of classifiers have been developed (most well known are support vector classifiers). In order to identify the decision boundary the available data are split into training and test data. The test data are put aside and only the training data are then used to find a decision boundary, as, for example, is shown in Figure 4e. The crucial test is then performed with the remaining test data. The classifier is applied to these data to see to which degree it is able to correctly assign the labels. Depending on which side of the decision boundary a test data point falls upon, it will yield either a correct or an incorrect classification.

Please note the similarity between the mapping of mental states and brain states. The red cloud of points in Figure 4b shows a two-dimensional response pattern that corresponds to the neural code for percepts of dogs. The spread of the point cloud (i.e. the fact that repeated measurements don’t yield identical values) could mean two things. Either the spread reflects *noise and uncertainty* that is typically inherent in measurements of neural data. This could, for example, reflect the fact that single fMRI voxels can sample many thousand cells, only few of which might be involved in processing. Additionally, physiological background rhythms can influence the signals and contribute to noise (Fox et al. 2006). Alternatively, however, the spread of the points could also be an inherent property of the representation. This would suggest that every time a person sees or visual imagines a dog, a slightly different activation pattern is observed in the brain. This would then

be evidence of *multiple realization*. Current measurement techniques do not have sufficient precision to distinguish between these two accounts. One difference between the multivariate mapping shown in Figure 3a (right) and the classification in Figure 4 is that the classification shows response distributions where each individual variable (voxel, channel) can adopt a graded value, whereas the values in Figure 3a (right) are only binary.

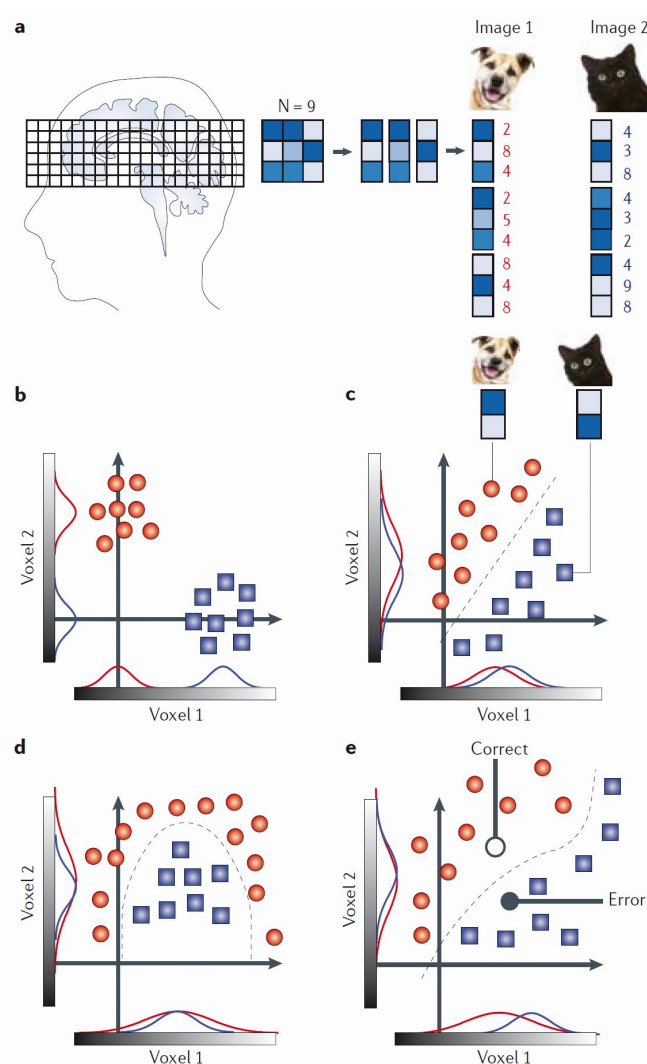


Figure 4: Mental state decoding using classification techniques (image adapted from Haynes & Rees 2005).

5 What does multivariate decoding reveal about NCCCs?

The importance of information theory for understanding the neural correlates of consciousness has been stressed repeatedly, most notably

by Giulio Tononi (2005). His information integration theory focuses on the information-based properties of neural processes and uses these special properties to provide a general explanation of consciousness. In contrast, the multivariate decoding account presented here attempts to solve the much more basic question of which neural populations provide the best account for which visual experiences. As mentioned above, this can be thought of as a search for the core neural correlates of the contents of consciousness (NCCCs), which have been postulated in similar forms by previous authors (Chalmers 2000; Block 2007; Koch 2004). While these proposals for core NCCCs have been influential in theoretical discussions on consciousness, they have only rarely been directly linked to neuroscience research, which requires spelling out how the NCCC can be established in empirical data. In the following, various studies of multivariate decoding from our lab will be presented that have implications for identifying NCCCs.

5.1 Example 1: Encapsulated information in V1

There has long been a debate as to whether the primary visual cortex (V1) is a neural correlate of visual consciousness. Crick & Koch (1995) postulated that V1 does not encode visual experiences for several reasons. First, V1 does not have the anatomical projections to the pre-frontal cortex that would allow for a direct read-out of information in V1. This would be required to explain a key distinguishing feature of conscious experiences: that we can voluntarily act upon them. A second reason is that V1 encodes information of which we are not aware. Psychophysical experiments, for example, show that V1 can encode orientation information of which we are not aware (He et al. 1996). We thus directly assessed the link between information encoding in V1 and visual awareness (Haynes & Rees 2005). Specifically, we investigated the effects crossing the threshold to awareness has on the neural coding of simple visual features. Participants viewed oriented “grating” images (Figure 5) and had to tell whether they were tilted to the left or to the right. In one

condition the images were clearly visible, in the other condition they were rendered invisible by rapidly alternating the orientation stimulus with a mask. In this condition participants were not able to tell the difference between the two orientation stimuli (Figure 5).

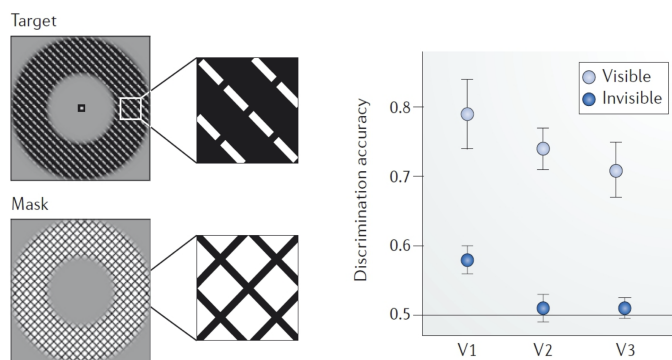


Figure 5: Decoding the orientation of invisible grating stimuli from patterns of activity in early visual areas. Target stimuli were line patterns that were either tilted top left to bottom right, or top right to bottom left. They were rapidly alternated with mask stimuli so that participants were unable to identify the target orientation. The classification accuracy for these “invisible” gratings was above chance in area V1, but not in V2 or V3. For visible orientation stimuli the classification was above chance in all three early visual areas (figure taken from Haynes & Rees 2005).

We then applied a classifier to fMRI-signals from early visual regions V1, V2, and V3 to see if it would be possible to decode the orientation of stimuli. We found that orientation for the visible stimuli could be decoded from all early visual regions, V1, V2, and V3 (Figure 5, right). This is in line with previous research on encoding of orientation information in early visual areas (Bartfeld & Grinvald 1992). Interestingly, we were able to decode the orientation from V1 even for invisible stimuli. This means that V1 presumably continues to carry low-level feature information even when a participant can’t access this information. V2 and V3, however, only had information for visible stimuli, not for invisible stimuli. Please note that an alternative interpretation could be that subjects perceive the subtle differences between masked stimuli, but they cannot report or reason about them. However, in psychophysics an absence of

discriminability is typically considered a strong criterion for absence of awareness. This finding is interesting for several reasons. First, it demonstrates that information can be encapsulated in a person’s early visual cortex, without them being able to access this information. This suggests that V1 is not an NCCC for conscious orientation perception. Second, it shows that one explanation why stimuli are rendered invisible by visual masking is that the information that is available at early stages of processing (V1) is not passed on to the next stages of processing in V2 and V3. Similar encapsulation of information has also been observed for parietal extinction patients in higher-level visual areas with more conventional neuroimaging approaches (Rees et al. 2000).

5.2 Example 2: Imagery and perception

There has also been a long debate on the neural mechanisms underlying visual imagery. One important question is whether the NCCCs underlying imagery are the same—or at least overlapping—with those for veridical perception. One study (Kosslyn et al. 1995) found that imagery activated even very early stages of the visual cortex. This fits with a mechanism that encodes visual images as a replay of representations of veridical percepts. However, this does not reveal whether the activation of the early visual cortex really participates in encoding the imagined contents. Instead, these regions might be involved in ensuring the correct spatial distribution of attention across the visual field (Tootell et al. 1998). The question of whether the neural representations of veridical percepts are the same as those for visual imaginations needs to be established in addition.

We conducted a study to directly address the overlap of NCCCs for veridical perception and imagery (Cichy et al. 2012). Participants were positioned inside an MRI scanner and had to perform one of two tasks: Either they were asked to *observe* visual images presented to the left or right of fixation (Figure 6), or they were asked to *imagine* visual images in the same locations. Twelve different images from four categories were used: three objects, three visual

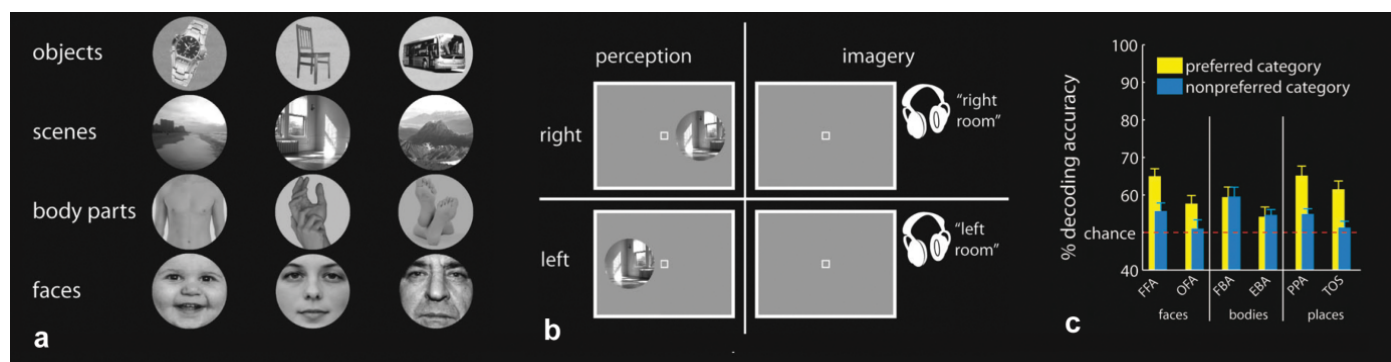


Figure 6: Visual imagery. (a) Visual stimuli used in the experiment consisted of three selections from four categories. (b) In different trials participants either saw the images to the left or right of fixation or they received an auditory instruction to imagine a specific image. (c) A classifier trained on the brain responses of different imagined images could be used able to correctly cross-classify which image a person was currently seeing on the screen in the perception condition. Information was higher for the images “preferred” by a visual area, but there was still information, esp. in FBA, about the non-preferred categories (FFA=fusiform face area; OFA=occipital face area; FBA=fusiform body area; EBA=extrastriate body area; PPA=parahippocampal place area; TOS=transverse occipital sulcus)(figure from Cichy et al. 2012).

scenes, three body parts, and three faces. We found that multiple higher-level visual regions had information about the images. Furthermore, it was possible to decode seen visual images using a classifier that had only been trained on imagined visual images. This suggests that imagery and veridical perception share *similar* neural representations for perceptual contents, at least in high-level visual regions. Please note, however, that the cross-classification between veridical perception and imagery is not perfect. It is currently unclear whether this reflects imperfections in the measurement of brain signals with fMRI, or whether it reflects residual differences in the contents of consciousness between imagery and veridical perception, for example the higher vividness of perception based on external visual stimuli (Perkey 1910).

5.3 Example 3: Perceptual learning

Another interesting riddle of sensory awareness is perceptual learning (Sagi 2011; see also Lamme this collection). When we are first exposed to a novel class of sensory stimuli our ability to differentiate between nuances is highly limited. When one tastes the first glass of wine, all wines taste the same. But with increased exposure and experience we learn to distinguish even subtle differences between different wines.

The interesting question here is whether the sensory information was there all along, and we just failed to notice it, or whether the sensory representation of the wines actually improves (see Dennett 1991).

We addressed this question, but with visual grating stimuli instead of different wines (Kahnt et al. 2011). Participants performed two fMRI sessions, where they had to distinguish small differences in the orientation of lines presented on the screen. They had to tell whether they were rotated clockwise or counter-clockwise with respect to a template. During the first fMRI session their ability to distinguish between the line patterns was quite limited. Afterwards we trained them in two sessions outside the MRI scanner on the same line patterns, and their performance continually improved. In a final second fMRI session they had then substantially improved their ability to tell even subtle differences between the orientations apart. But what explains this improvement: Better sensory coding, or better interpretation of the information that was there all along?

To address this question we first looked into the responses in the early visual cortex to the different line stimuli. As expected from our above-mentioned study on orientation coding (Haynes & Rees 2005), it was possible to decode the orientation of the line elements from signals in early

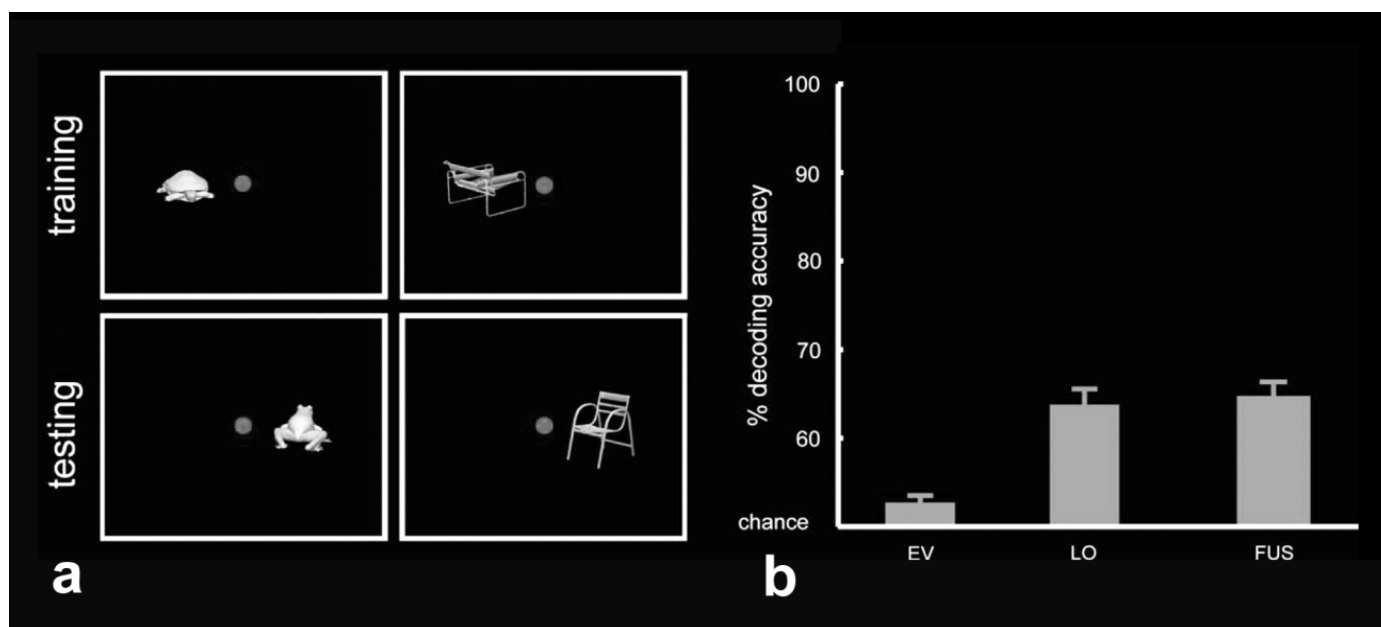


Figure 7: FMRI evidence for invariance of object-representations in the high-level visual regions lateral occipital (LO) and fusiform gyrus (FUS) as compared to early visual cortex (EV; figure from [Cichy et al. 2011](#)).

visual areas. It is well established that these areas have information about such simple visual features ([Bartfeld & Grinvald 1992](#)). However, we found no improvement in our ability to decode the orientation of the stimuli with learning. There is some divergence in the literature with some studies finding effects of learning in early sensory areas (see [Sasaki et al. 2010](#)). Other recent findings in monkeys are in line with our findings and do not find improved information coding in sensory areas (e.g., [Law & Gold 2009](#)). In our case, it seems as if the sensory representation of orientation remains unchanged and that some other mechanism has to be responsible for the improvement in perceptual discrimination. We found a region in the medial prefrontal cortex where signals followed the learning curve, thus suggesting that the improvement was not so much a question of stimulus coding but of the *read-out* of information from the sensory system. This study suggests that representation of a feature in an NCCC might not automatically guarantee it enters visual awareness.

5.4 Example 4: Invariance in human higher-level visual cortex

As mentioned above, one important challenge to the idea that the contents of visual awareness

are encoded exclusively late in the visual system is the invariance of responses to low-level visual features ([Sáry et al. 1993](#)). We directly investigated the invariance of fMRI responses in the regions lateral occipital (LO) and fusiform gyrus (FUS) of the higher-level object-selective visual cortex ([Malach et al. 1995](#); [Grill-Spector et al. 2001](#)). In this study ([Cichy et al. 2011](#)) participants viewed objects presented either to the left or the right of the fixation spot ([Figure 7](#)). These objects consisted of three different exemplars from four different categories (animals, planes, cars, and chairs). For example, the category “animal” contained images of a frog, a tortoise, and a cow. With these data we were able to explore two different aspects of invariance. First, we wanted to know whether object representations are invariant to changes in spatial location. This is important because a low-level visual representation that focuses exclusively on the distribution of light in the visual field would not be able to generalize from one position to another. So we assessed whether a classifier trained to recognize an object at one position in the visual field would be able to generalize to another position in the visual field. We found that a classifier was able to generalize to a different position, however with reduced accuracy. This indicates that the representations

were at least partially invariant with respect to low-level visual features. Next, we investigated whether the representations would generalize from one exemplar to another. This goes even further in testing for the level of abstraction of the representation. A classifier that can generalize not only to a different location but even to a different exemplar (say from a frog to a cow) needs to operate at a higher level of abstraction that is largely independent from low-level visual features. Again we found that the classifier was able to generalize between exemplars of the same category, further supporting the abstraction of representations in the higher visual regions LO and FUS (Figure 7). This makes it again less plausible that the contents of visual awareness are encoded exclusively in the higher-level visual cortex. Encoding in these regions is invariant (or at least tolerant) to low-level feature changes, and thus this level of perceptual experience has to be encoded at a different, presumably lower, level of visual processing.

5.5 Example 5: No sensory information in PFC

A further case where multivariate decoding might inform theories of visual awareness becomes apparent when we confront the question of whether sensory information is distributed throughout the brain when a stimulus crosses the threshold of awareness. The global neuronal workspace theory (e.g., Dehaene & Naccache 2001; see also Baars 2002) posits that sensory signals are made globally *available* across large-scale brain networks, especially in the prefrontal and parietal cortices, when they reach awareness. An interesting and open question is whether this global availability of sensory information means that the sensory information about a stimulus can be actually decoded from these prefrontal and parietal brain regions to which the information is made available. In theory, one might be able to distinguish between a “streaming model” of global availability, where information is broadcast throughout the brain (e.g., Baars 1988), and which should thus be decodable from multiple brain regions; an alternative would be an “on demand” model of global

availability, where sensory signals are only propagated into prefrontal and parietal cortex when selected by attention (e.g., Dehaene & Naccache 2001).

We performed three fMRI studies to test this question (Bode et al. 2012; Bode et al. 2013; Hebart et al. 2012). In the first study (Bode et al. 2012), participants were briefly shown images of pianos and chairs that were temporally embedded in scrambled mask stimuli. There were two conditions. In one condition, the timing of visual stimuli was chosen such that the target stimuli were clearly visible. In the other condition, the timing of scrambled masks and targets was such that the targets were effectively rendered invisible. We attempted to decode the sensory information about the presented objects. Under high visibility we were able to decode which image was being shown from fMRI signals in the so-called lateral occipital regions of the human brain, where complex object recognition takes place. Under low visibility, there was no information in these brain regions. This suggests a possible mechanism for explaining why the stimuli failed to reach awareness. Presumably their sensory representations were already cancelled out at the visual processing stages. The “streaming model” mentioned above would mean that sensory information about the object category is distributed into parietal and prefrontal brain regions when the stimulus crosses the threshold of awareness. However, we found no information in the prefrontal cortex—under either high or low visibility (Bode et al. 2012). This finding was repeated in two different studies, one also using objects as stimuli (Bode et al. 2013) and one using drifting motion stimuli (Hebart et al. 2012). In contrast, in animal studies sensory information has been found in the prefrontal cortex (Pasternak & Greenlee 2005). It is currently unclear whether this reflects a species-difference or whether it is due to limitations in the resolution of human neuroimaging techniques.

5.6 Example 6: Unconscious processing of preferences

It is well known that unattended and even invisible visual stimuli can undergo substantial processing. We investigated whether informa-

tion about high-level, more interpretative and subjective properties of visual stimuli would also be traceable using decoding techniques. For this we aimed to decode the degree to which *preferences* for certain visually presented images of cars can be decoded, even when these stimuli were unattended and were not task-relevant (Tusche et al. 2010).

For this experiment we carefully pre-selected our participants, who were self-reporting car-enthusiasts. Then we ensured that we chose stimuli where different participants had maximally-divergent opinions as to which car they preferred. This was necessary in order to de-correlate the classification of the preference from the classification of the specific vehicles. Subjects were divided into two groups. Participants from the first group were presented with the car images in the scanner and had to actively evaluate whether they liked them. The second group was also presented with the car images, but they were distracted from them. They were required to solve a very difficult task that required them to focus their attention elsewhere in the visual field, on fixation. The car stimuli were thus task-irrelevant and presented outside of the attentional focus. This group of subjects could not recall which cars had been shown during the experiment, suggesting that they were indeed not actively deliberating about the cars. After the experiment, participants from both groups were asked to rate how much they would like to buy each car. This served as a gold standard for their preference.

We then tried to decode whether individual subjects liked the cars or not. For this, we looked into patterns of brain activity throughout the brain, to see where there might be information regarding preferences. This was done in order to reduce the bias when only looking into pre-specified brain regions. We found that it was possible to decode the preferred cars with 75% accuracy from brain regions far outside the visual system, in the medial prefrontal and in the insular cortex. This was true for the subjects who had been actively deliberating about their preferences for the cars, but also for the participants who been distracted from thinking about them. Presumably, this

means that the brain automatically processes the car-images all the way up to the stage of encoding preferences, even in the absence of visual attention. Please note that this finding of *preference* information in the prefrontal cortex is quite different to that in the previous experiment, where there was no *sensory* information in PFC. Here, in contrast, there is information in PFC, but (a) not about a sensory property and (b) even for unattended stimuli. Thus, it appears that the informational dividing line between sensory and prefrontal brain regions is not one of awareness, but rather one of the type of information coded.

6 Complicating factors

When searching for the neural correlates of contents of consciousness (NCCCs), there are several complicating factors. One might a priori have an assumption of modularity, meaning that one feature is encoded in one dedicated NCCC area. The idea that such single, maximally-informative regions exist for different features, however, is no more than an assumption. It might turn out that perceptual coding—even for single features—inherently involves processes in multiple brain regions.

Empirically, it is known that information about objects is distributed across multiple regions. One question is whether one brain region can have information about more than one content (e.g., Haxby et al. 2001; Cichy et al. 2013). In a study inspired by Haxby et al. (2001) we investigated whether object-selective brain regions have information only about objects from their preferred category (Cichy et al. 2013). Participants viewed images from four different categories: objects, visual scenes, body parts, and faces. These categories were chosen because faces, body parts, and places are believed to be processed by highly selective brain regions. We found that a classifier not only contained information about a region's preferred category: take the example of the face-selective region FFA. It was not only possible to classify faces from this region, it was also possible to classify the difference between other, non-face-related objects, say between a chair and a window. The

flipside of this finding that individual regions encode multiple contents is that individual perceptual contents were found in multiple regions. For example, information about faces was also found in supposedly “non-face-selective” brain regions (e.g., in the PPA). This presents a challenge to the idea that each content is represented in one region only.

However, the problem might not be as severe as it first appears. It is actually expected that multiple regions will contain information about each type of content. Different brain regions do not exist in isolation, but are densely causally interconnected (Felleman & van Essen 1991). Furthermore, in the visual pathway, stimulus-related information will reach higher-level brain regions by way of low-level regions. Even if the FFA is the visual region that responds most (albeit not fully) selectively to faces, the presence of a face could also be inferred from the discharge pattern of ganglion cells in the retina. Thus, vertical and horizontal propagation of information is expected. One crucial criterion, which has not received much attention, is whether the information in different regions is *redundant* or whether it is *independent* with respect to a person’s perceptual experience. If one hypothetical brain region, say the uniform unicorn area (UUA), is directly responsible for visual experiences of unicorns, it should have more information about a person’s unicorn experiences than any other region.

The relationship between information in the UUA and other areas will reveal a lot about the nature of representation. If other regions also have information about unicorn experiences, and they receive their information about unicorn experiences via the UUA, then the unicorn-related information in the other regions should be partially redundant to that in the UUA. A classifier should not be able to extract more information about a person’s unicorn experiences by additionally taking other regions into account, over and above the information available from the UUA. If, in contrast, other regions have information that goes beyond that in the UUA that allow the system to improve the classification of unicorn experiences, it is likely that the representation itself is distrib-

uted across multiple brain regions. Another way to put it is to distinguish between representational and causal entanglement. A change in neural activity in one region will typically be propagated to any neighbouring regions with which it is connected. This *causal* entanglement, however, does not directly implicate *representational* entanglement. Only if it were not possible to find an individual region where neural activity patterns is not fully informative of a specific feature, and if taking into account the joint activity of this region and another region did provide full information, would this provide evidence for representational entanglement.

7 Putting it together

As outlined above, when attempting to identify the neural correlate of a particular content of conscious experience, it is important to ensure that brain representations in any candidate region fulfil certain mapping requirements. Because we have no direct way of establishing this mapping, multivariate decoding provides a rough approximation that allows the linking of perceptual contents to population brain responses in different regions, and allows us to explore their properties. The data from our lab provide several constraints for a theory of NCCs. Consistent with previous suggestions (Crick & Koch 1995), the very early stages of processing in V1 are presumably not directly involved in encoding visual experiences. Representations in these regions have more detail than enters consciousness (Haynes & Rees 2005) and might not change their information content during perceptual learning when contents are successively represented with more detail in consciousness (Kahnt et al. 2011). Please note that early regions beyond V1 have to be NCCs, because higher-level visual areas are invariant to low-level visual features. This has not only been shown in animals (Sáry et al. 1993), but also in humans using classification techniques (e.g., Cichy et al. 2011). This invariance means that high-level regions cannot simultaneously encode the high-level, more abstract phenomenal properties (such as whether a cloud of points resembles a dog or a cat) and the low-level phe-

nominal properties (colour or brightness sensations). Multiple regions are needed to account for the full multilevel nature of our perceptual experience. While V1 is presumably excluded from visual awareness, early extrastriate regions (such as V2) are likely to be involved, because they still encode low-level visual information. They also appear to filter out sensory information that does not enter awareness, thus again closely matching perceptual experience. For example, V2 and V3 do not encode the orientation of invisible lines, whereas V1 does (Haynes & Rees 2005). Similarly, neural object representations in the lateral occipital complex were wiped out by visual masking that rendered an object stimulus invisible (Bode et al. 2012). The role of extrastriate and higher-level visual areas in visual awareness is further highlighted by the fact that they exhibit a certain convergence of different aspects of awareness. Most notably, they employ a shared code for visual perception and visual imagery (Cichy et al. 2012).

While extrastriate and higher-level visual regions jointly encode different feature levels of visual awareness, there is evidence that a representation in these regions is not sufficient for visual awareness. For example, our experiments on perceptual learning (Kahnt et al. 2011)—where subjects are unable to access certain details of visual stimuli—show that improved sensory perception is not necessarily associated with improved representation of information in these early areas. The mechanism through which perception of details might be improved lies beyond the sensory encoding stage, in the prefrontal cortex. The mechanism of this improvement is not an improved sensory representation in the prefrontal cortex. Contrary to several experiments on animals (Pasternak & Greenlee 2005), our experiments consistently fail to show any sensory information in the frontal cortex. For example, when a stimulus survives visual masking and is consciously perceived, there is no evidence for the additional distribution of information into the prefrontal cortex (Bode et al. 2012; Bode et al. 2013; Hebart et al. 2012) as would be expected if information is indeed made globally available in the sense of a “streaming model” of a global

workspace (Dehaene & Naccache 2001). Even in a more conventional experimental task, based on visual working memory, we were not able to identify sensory information in the prefrontal cortex. Thus, the direct encoding of the visual contents of consciousness, the NCCCs appear to lie in sensory brain regions, at least as far as can be told with the resolution of non-invasive human neuroimaging techniques. On the other hand, our results suggest that the prefrontal cortex is involved in the decision-making—as has been suggested before (Heekeren et al. 2004)—and in learning about sensory contents (Kahnt et al. 2011). Thus, it appears to do so without re-representing or encoding sensory information itself.

References

- Anders, S., Heinzle, J., Weiskopf, N., Ethofer, T. & Haynes, J. D. (2011). Flow of affective information between communicating brains. *NeuroImage*, 54 (1), 439-446. [10.1016/j.neuroimage.2010.07.004](https://doi.org/10.1016/j.neuroimage.2010.07.004)
- Azevedo, F. A. C., Carvalho, L. R. B., Grinberg, L. T., Farfel, J. M., Ferretti, R. E. L., Leite, R. E. P., Filho, W. S., Lent, R. & Herculano-Houzel, S. (2009). Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *Journal of Comparative Neurology*, 513 (5), 532-541. [10.1002/cne.21974](https://doi.org/10.1002/cne.21974)
- Baars, B. J. (1988). *A cognitive theory of consciousness*. Cambridge, UK: Cambridge University Press.
- (2002). The conscious access hypothesis: origins and recent evidence. *Trends in Cognitive Sciences*, 6 (1), 47-52. [10.1016/s1364-6613\(00\)01819-2](https://doi.org/10.1016/s1364-6613(00)01819-2)
- Bartfeld, E. & Grinvald, A. (1992). Relationships between orientation-preference pinwheels, cytochrome oxidase blobs, and ocular-dominance columns in primate striate cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 89 (24), 11905-11909. [10.1073/pnas.89.24.11905](https://doi.org/10.1073/pnas.89.24.11905)
- Block, N. (2007). Consciousness, accessibility and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences*, 30 (5-6), 481-499. [10.1017/S0140525X07002786](https://doi.org/10.1017/S0140525X07002786)
- Bode, S., Bogler, C., Soon, C. S. & Haynes, J. D. (2012). The neural encoding of guesses in the human brain. *NeuroImage*, 59 (2), 1924-1931. [10.1016/j.neuroimage.2011.08.106](https://doi.org/10.1016/j.neuroimage.2011.08.106)
- Bode, S., Bogler, C. & Haynes, J. D. (2013). Similar neural mechanisms for perceptual guesses and free decisions. *NeuroImage*, 65, 456-465. [10.1016/j.neuroimage.2012.09.064](https://doi.org/10.1016/j.neuroimage.2012.09.064)
- Chalmers, D. (2000). What is a neural correlate of consciousness? In T. Metzinger (Ed.) *Neural Correlates of Consciousness: Conceptual and Empirical Questions* (pp. 17-40). Cambridge, MA: MIT Press.
- Cichy, R. M., Chen, Y. & Haynes, J. D. (2011). Encoding the identity and location of objects in human LOC. *NeuroImage*, 54 (3), 2297-2307. [10.1016/j.neuroimage.2010.09.044](https://doi.org/10.1016/j.neuroimage.2010.09.044)
- Cichy, R. M., Heinzle, J. & Haynes, J. D. (2012). Imagery and perception share cortical representations of content and location. *Cerebral Cortex*, 22 (2), 372-380. [10.1093/cercor/bhr106](https://doi.org/10.1093/cercor/bhr106)
- Cichy, R. M., Sterzer, P., Heinzle, J., Elliot, L. T., Ramirez, F. & Haynes, J. D. (2013). Probing principles of large-scale object representation: category preference and location encoding. *Human Brain Mapping*, 34 (7), 1636-1651. [10.1002/hbm.22020](https://doi.org/10.1002/hbm.22020)
- Crick, F. & Koch, C. (1995). Are we aware of neural activity in primary visual cortex? *Nature*, 375 (6527), 121-123. [10.1038/375121a0](https://doi.org/10.1038/375121a0)
- Dehaene, S. & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, 79 (1-2), 1-37. [10.1016/S0010-0277\(00\)00123-2](https://doi.org/10.1016/S0010-0277(00)00123-2)
- Dennett, D. (1991). *Consciousness explained*. Boston, MA: Penguin.
- Engel, A. K. & Singer, W. (2001). Temporal binding and the neural correlates of sensory awareness. *Trends in Cognitive Sciences*, 5 (1), 16-25. [10.1016/S1364-6613\(00\)01568-0](https://doi.org/10.1016/S1364-6613(00)01568-0)
- Felleman, D. J. & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1 (1), 1-47. [10.1093/cercor/1.1.1](https://doi.org/10.1093/cercor/1.1.1)
- Formisano, E., Kim, D. S., Di Salle, F., van de Moortele, P. F., Ugurbil, K. & Goebel, R. (2003). Mirror-symmetric tonotopic maps in human primary auditory cortex. *Neuron*, 40 (4), 859-869. [10.1016/S0896-6273\(03\)00669-X](https://doi.org/10.1016/S0896-6273(03)00669-X)
- Fox, M. D., Snyder, A. Z., Zacks, J. M. & Raichle, M. E. (2006). Coherent spontaneous activity accounts for trial-to-trial variability in human evoked brain responses. *Nature Neuroscience*, 9 (1), 23-25. [10.1038/nn1616](https://doi.org/10.1038/nn1616)
- Fujita, I., Tanaka, K., Ito, M. & Cheng, K. (1992). Columns for visual features of objects in monkey inferotemporal cortex. *Nature*, 360 (6402), 343-346. [10.1038/360343a0](https://doi.org/10.1038/360343a0)
- Grill-Spector, K., Kourtzi, Z. & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research*, 41 (10-11), 1409-1422. [10.1016/S0042-6989\(01\)00073-6](https://doi.org/10.1016/S0042-6989(01)00073-6)
- Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L. & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293 (5539), 2425-2430. [10.1126/science.1063736](https://doi.org/10.1126/science.1063736)
- Haynes, J. D. (2009). Decoding visual consciousness from human brain signals. *Trends in Cognitive Sciences*, 13 (5), 194-202. [10.1016/j.tics.2009.02.004](https://doi.org/10.1016/j.tics.2009.02.004)
- Haynes, J. D. & Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary

- visual cortex. *Nature Neuroscience*, 8 (5), 686-691. [10.1038/nm1445](#)
- (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7 (7), 523-534. [10.1038/nrn193](#)
- Hebart, M. N., Donner, T. H. & Haynes, J. D. (2012). Human visual and parietal cortex encode visual choices independent of motor plans. *NeuroImage*, 63 (3), 1393-1403. [10.1016/j.neuroimage.2012.08.027](#)
- Heekeren, H. R., Marrett, S., Bandettini, P. A. & Ungerleider, L. G. (2004). A general mechanism for perceptual decision-making in the human brain. *Nature*, 431 (7010), 859-862. [10.1038/nature02966](#)
- He, S., Cavanagh, P. & Intriligator, J. (1996). Attentional resolution and the locus of visual awareness. *Nature*, 383 (6598), 334-337. [10.1038/383334a0](#)
- Kahnt, T., Grueschow, M., Speck, O. & Haynes, J. D. (2011). Perceptual learning and decision-making in human medial frontal cortex. *Neuron*, 70 (3), 549-559. [10.1016/j.neuron.2011.02.054](#)
- Koch, C. (2004). *The quest for consciousness: A neurobiological approach*. Englewood, CL: Roberts & Company.
- Kosslyn, S. M., Thompson, W. L., Kim, I. J. & Alpert, N. M. (1995). Topographical representations of mental images in primary visual cortex. *Nature*, 378 (6556), 496-498. [10.1038/378496a0](#)
- Lamme, V. A. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*, 10 (11), 494-501. [10.1016/j.tics.2006.09.001](#)
- (2015). The crack of dawn: Perceptual functions and neural mechanisms that mark the transition from unconscious processing to conscious vision. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Law, C. T. & Gold, J. I. (2009). Reinforcement learning can account for associative and perceptual learning on a visual-decision task. *Nature Neuroscience*, 12 (5), 655-663. [10.1038/nrn.2304](#)
- Leopold, D. A. & Logothetis, N. K. (1996). Logothetis N. K. (1996). Activity changes in early visual cortex reflect monkeys' percepts during binocular rivalry. *Nature*, 379 (6565), 549-553.
- Malach, R., Reppas, J. B., Benson, R. R., Kwong, K. K., Jiang, H., Kennedy, W. A., Ledden, P. J., Brady, T. J., Rosen, B. R. & Tootell, R. B. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 92 (18), 8135-8139.
- Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M. A., Morito, Y., Tanabe, H. C., Sadato, N. & Kamitani, Y. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60 (5), 915-925. [10.1016/j.neuron.2008.11.004](#)
- Müller, K. R., Mika, S., Rätsch, G., Tsuda, K. & Schölkopf, B. (2001). An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12 (2), 181-201. [10.1109/72.914517](#)
- Pascual-Leone, A. & Walsh, V. (2001). Fast backprojections from the motion to the primary visual area necessary for visual awareness. *Science*, 292 (5516), 510-512. [10.1126/science.1057099](#)
- Pasternak, T. & Greenlee, M. W. (2005). Working memory in primate sensory systems. *Nature Reviews Neuroscience*, 6 (3), 97-107. [10.1038/nrn1637](#)
- Penfield, W. & Rasmussen, T. (1950). *The cerebral cortex of man*. New York, NY: Macmillan.
- Perkey, C. (1910). An experimental study of imagination. *American Journal of Psychology*, 21 (3), 422-452. [10.1037/h0041622](#)
- Quiroga, R. Q., Kreiman, G., Koch, C. & Fried, I. (2008). Sparse but not 'grandmother-cell' coding in the medial temporal lobe. *Trends in Cognitive Sciences*, 23 (3), 87-91. [10.1016/j.tics.2007.12.003](#)
- Raffman, D. (1995). On the persistence of phenomenology. In T. Metzinger (Ed.) *Conscious experience* (pp. 293-308). Paderborn, GER: Schöningh Verlag.
- Rees, G., Wojciulik, E., Clarke, K., Husain, M., Frith, C. & Driver, J. (2000). Unconscious activation of visual cortex in the damaged right hemisphere of a parietal patient with extinction. *Brain*, 23 (8), 1624-1633. [10.1093/brain/123.8.1624](#)
- Sagi, D. (2011). Perceptual learning in vision research. *Vision Research*, 51 (13), 1552-1566. [10.1016/j.visres.2010.10.019](#)
- Sasaki, Y., Nanez, J. E. & Watanabe, T. (2010). Advances in visual perceptual learning and plasticity. *Nature Reviews Neuroscience*, 11 (1), 53-60. [10.1038/nrn2737](#)
- Sereno, M. I., Dale, A. M., Reppas, J. B., Kwong, K. K., Belliveau, J. W., Brady, T. J., Rosen, B. R. & Tootell, R. B. (1995). Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science*, 268 (5212), 889-893. [10.1126/science.7754376](#)
- Sheinberg, D. L. & Logothetis, N. K. (1997). The role of temporal cortical areas in perceptual organization.

- Proceedings of the National Academy of Sciences of the United States of America*, 94 (7), 3408-3413.
- Singer, W. (2015). The ongoing search for the neuronal correlate of consciousness. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Sáry, G., Vogels, R. & Orban, G. A. (1993). Cue-invariant shape selectivity of macaque inferior temporal neurons. *Science*, 260 (5110), 995-997.
- Tong, F., Nakayama, K., Vaughan, J. T. & Kanwisher, N. (1998). Binocular rivalry and visual awareness in human extrastriate cortex. *Neuron*, 21 (4), 753-759.
[10.1016/S0896-6273\(00\)80592-9](https://doi.org/10.1016/S0896-6273(00)80592-9)
- Tononi, G. (2005). Consciousness, information integration, and the brain. *Progress in Brain Research*, 150, 109-126.
- Tononi, G., Srinivasan, R., Russell, D. P. & Edelman, G. M. (1998). Investigating neural correlates of conscious perception by frequency-tagged neuromagnetic responses. *Proceedings of the National Academy of Sciences of the United States of America*, 95 (6), 3198-3203.
- Tootell, R. B., Hadjikhani, N., Hall, E. K., Marrett, S., Vanduffel, W., Vaughan, J. T. & Dale, A. M. (1998). The retinotopy of visual spatial attention. *Neuron*, 21 (6), 1409-1422.
[10.1002/\(SICI\)1097-0193\(1997\)5:4<280::AID-HBM13>3.0.CO;2-I](https://doi.org/10.1002/(SICI)1097-0193(1997)5:4<280::AID-HBM13>3.0.CO;2-I)
- Tusche, A., Bode, S. & Haynes, J. D. (2010). Neural responses to unattended products predict later consumer choices. *Journal of Neuroscience*, 30 (23), 8024-8031.
[10.1523/JNEUROSCI.0064-10.2010](https://doi.org/10.1523/JNEUROSCI.0064-10.2010)
- Yacoub, E., Harel, N. & Ugurbil, K. (2008). High-field fMRI unveils orientation columns in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 105 (30), 10607-10612.
[10.1073/pnas.0804110105](https://doi.org/10.1073/pnas.0804110105)

What's up with Prefrontal Cortex?

A Commentary on John-Dylan Haynes

Caspar M. Schwiedrzik

The prefrontal cortex is perhaps one of the most intriguing areas of the brain, and considered by many to be involved in a whole battery of higher cognitive functions. However, evidence for a direct involvement in conscious perception, although often postulated, remains inconclusive. In his paper, John-Dylan Haynes presents results from experiments using multivariate decoding techniques on human functional magnetic resonance imaging data that speak against the assertion that prefrontal cortex broadcasts the contents of consciousness throughout the brain. I consider potential reasons for these null results, as well as where else we may look for the neural correlates of consciousness. Specifically, I propose that conscious perception arises when distributed neurons are bound into coherent assemblies—a process that does not require relay through specific brain areas.

Keywords

Multivariate pattern analysis | Neuronal correlates of consciousness | Neuronal synchrony | Prefrontal cortex

Commentator

Caspar Schwiedrzik

cschwiedrz@rockefeller.edu

The Rockefeller University
New York, NY, U.S.A.

Target Author

John-Dylan Haynes

haynes@bccn-berlin.de

Charité – Universitätsmedizin Berlin
Berlin, Germany

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

There is a striking parallel between the hierarchical organization of behavior and the hierarchical organization of the cerebral cortex (Botvinick 2008). It is thus tempting to assign consciousness, at least historically often considered to be one of our highest functions (Jackendoff 1987; Markowitsch 1995), to the prefrontal cortex (PFC), which is positioned at the top of the cortical hierarchy. While the idea that consciousness can be localized to a single brain area has now been discredited, many current theories of consciousness still consider the PFC a key player in the emergence of conscious perception (Dehaene & Changeux 2011; Lau & Rosenthal 2011). And indeed, a multitude of

neuroimaging studies has shown differential activation for perceived vs. unperceived stimuli in various parts of the PFC (Dehaene et al. 2001; Lau & Passingham 2006; Sahraie et al. 1997; Schwiedrzik et al. 2014). A very prominent theoretical proposal on the neural correlates of consciousness, the Global Neuronal Workspace (GNW) model by Stanislas Dehaene and colleagues, posits that the PFC (in conjunction with parietal cortex) serves to distribute information that is processed in unconscious modules to the entire brain, and that it is this broadcasting of information that gives rise to conscious experience (Dehaene & Changeux 2011). The PFC may be particularly well equipped to

do so, for example because it hosts an abundance of neurons with long-distance connections, so called “von Economo” neurons, which seem ideally suited to both receive and deliver information from all areas of the brain to all areas of the brain (Dehaene & Changeux 2011). A prediction that can be directly derived from this account and that has been eloquently put forward by John-Dylan Haynes is that the PFC should at least temporarily represent the information that we consciously perceive, i.e., it should directly encode the contents of consciousness (Haynes 2009; this collection). To test this idea, Haynes and his coworkers have used a neuroimaging technique that allows for exquisite access to perceptual content, namely multivariate pattern analysis (MVPA) of functional magnetic resonance imaging (fMRI) signals. In this technique, powerful machine learning algorithms are used to analyze spatially-distributed patterns of brain activity, and a brain region is said to represent the content of interest if its activity patterns allow the reliable classification—in the case of consciousness—of which stimulus the subject perceived on a given trial. This contrasts with previous fMRI studies not using MVPA: because these studies do not directly address content, activity in the PFC (and other regions) that differentiates perceived from not perceived trials could in principle reflect other aspects of conscious experience, for example the allocation of attentional resources or working memory. The stunning result of Haynes’ investigations is that while MVPA shows that perceptual content can be decoded from higher sensory areas, PFC activity does not yield decoding accuracies higher than chance level. So what’s up with PFC?

2 Neuronal representations in the PFC

Indeed, the inability to decode perceptual content from the PFC runs counter intuitions about PFC functions we have from animal models such as the macaque monkey, where representations of (perceptual) content can be even more directly assessed than with MVPA, by using electrophysiological recordings from single neurons. These studies show that PFC neurons

are tuned for and thus represent perceptual features such as visual motion direction (Zaksas & Pasternak 2006) or somatosensory flutter frequency (Romo et al. 1999). Even more direct evidence for the representation of perceptual content in the PFC comes from a recent study by Theofanis Panagiotaropoulos et al. (2012), which shows that single PFC neurons exhibit stimulus-specific activity modulations as a function of subjective perception under flash suppression, a technique that can render visual stimuli temporarily invisible.

In the absence of direct electrophysiological recordings from human PFC, one possible explanation for this discrepancy is that the macaque brain is organized in a totally different way to the human brain. But while theoretically possible, this seems highly unlikely (Passingham 2009; Roelfsema & Treue 2014). Alternatively, one may consider whether certain properties of the neural representations in the PFC may pose limitations to the ability of the fMRI MVPA technique to decode their content. This is in light of the fact that decoding of content from human PFC has been unsuccessful not only in the context of conscious perception, but also in the context of working memory, which has led to a radical reinterpretation of the role of the PFC in this domain (Sreenivasan et al. 2014).

It has been hypothesized that successful decoding of stimulus features such as orientation or motion direction from sensory areas relies upon the presence of orderly spatial arrangements of these features in cortical columns or maps (Freeman et al. 2011; Kamitani & Tong 2005). It is thus worth asking the question whether PFC exhibits a similar map-like structure, or whether the spatial arrangement of features in the PFC already renders the likelihood of decoding any kind of information from its fMRI activity unlikely. For example, while maps representing space have been identified in the human PFC, they are much smaller than retinotopic maps in early visual areas, and intersubject variability in their location is much higher (Hagler & Sereno 2006). Furthermore, it is known from experiments in monkeys that only a subset of the neurons within the PFC subregions in which these maps have been found ac-

tually displays any spatial preference (Funahashi et al. 1989; Rainer et al. 1998). Importantly, the PFC also has a more complicated cytoarchitecture than sensory areas, with longer and more complex dendrites that allow for sampling of information from a wider range of inputs (Jacobs et al. 2001), which may affect the spatial scale at which information is represented and can be read out. Nevertheless, the overall picture that arises from studies employing optical imaging and microstimulation in monkeys is that at least several subregions of the PFC are topographically organized in a similar fashion as sensory areas (Roe 2010). However, while the topography of the PFC may be favorable to MVPA, neural representations in the PFC seem to exhibit more complex features and dynamics on a single neuron and population level than the representations in sensory areas where MVPA has been particularly successful. For example, recent studies in monkeys show that PFC representations are very high dimensional (Rigotti et al. 2013), that selectivity is not fixed but can be acquired (Bichot et al. 1996), that selectivity can change over time even within a trial (Stokes et al. 2013), and that populations of PFC neurons represent multiple stimulus dimensions at the same time even if one dimension is unattended (Mante et al. 2013). Thus, the dimensionality and temporal instability of neural representations in the PFC may pose a serious challenge to fMRI MVPA experiments, given that they rely on an inherently slow, hemodynamic signal that integrates neural activity over time.

Putting these and other (Anderson & Oates 2010; Vilarroya 2013) potential limitations of the MVPA approach aside, what other evidence do we have that the PFC is actually involved in conscious perception? In particular, is there *causal* evidence for a role of the PFC in conscious perception?

3 Beyond decoding: Causal evidence for a role of the PFC in conscious perception?

Early studies in macaque monkeys have found that lesions to the PFC can increase the lumin-

ance threshold (Latto & Cowey 1971) and degrade detection performance (Kamback 1973). More recently, studies in humans using transcranial magnetic stimulation have similarly found that stimulation of the PFC can impair the visibility of stimuli (Rounis et al. 2010), but also improve detection rates during visual masking (Grosbras & Paus 2003). Finally, Antoine Del Cul et al. have shown that perceptual thresholds are increased in patients with relatively small prefrontal lesions even when attentional effects are tightly controlled for (2009). However, none of these studies has shown dramatic impairments, but rather modulations of performance or perception. Total blindness has only been reported after removing the entire frontal cortex including (parts of) the underlying cingulate cortex in monkeys, and only lasted for a few days in several cases (Nakamura & Mishkin 1986). Importantly, other lesion studies in humans have not reported perceptual deficits at all (Heath et al. 1949; Markowitsch & Kessler 2000). Taken together with the fact that the PFC is also active during unconscious processing (Diaz & McCarthy 2007; Lau & Passingham 2007; van Gaal et al. 2008), not deactivated under Thiopental anesthesia (Veselis et al. 2004), but deactivated during rapid eye movement sleep when vivid (non-lucid) dreams can be experienced (Braun et al. 1998; Desseilles et al. 2011; Muzur 2002), this indicates that the evidence for a direct, specific involvement of the PFC in conscious perception is currently inconclusive at best.

4 An alternative to localization

Luckily, we can do without PFC, at least for the purposes of explaining conscious perception, while still maintaining many of the other, more compelling aspects of the GNW model. One central component that the GNW model shares with several other proposals about the neural correlates of consciousness (for example Melloni & Singer 2010; Tononi 2004; von der Malsburg 1997) is the concept of global integration of information. In light of the modular organization of the brain, a mechanism is required that brings information together such that an integrated, coherent percept

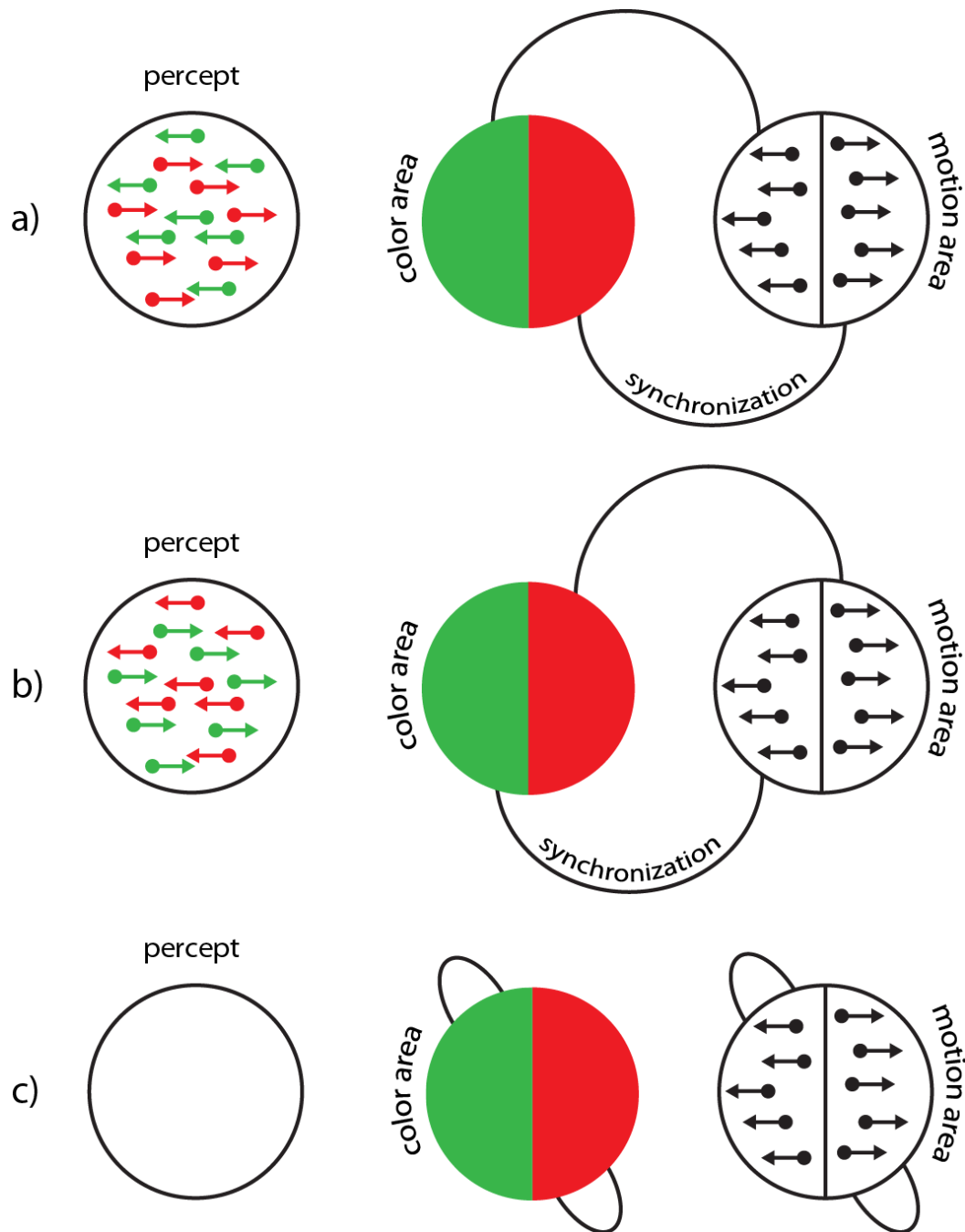


Figure 1: Neuronal synchrony binds distributed neurons into coherent assemblies, giving rise to conscious experience. Consider an experiment in which the subject is confronted with two superimposed, transparent surfaces of moving dots, as shown in the first column. **(a)** The dots on one surface are green and move to the left, and the dots on the other surface are red and move to the right. The two colors of dots are represented in a brain area coding for color, while the two motion directions are represented in an area coding for motion. If the neurons coding for green in the color area synchronize with the neurons coding for motion to the left in the motion area, and the neurons coding for red synchronize with the neurons coding for motion to the right, the two surfaces are consciously perceived. **(b)** A change in experience does not require a change in activity levels within areas, but a change of which neurons are synchronized. The opposite percept of (a) arises if neurons coding for green are synchronized with neurons coding for motion to the right, and if neurons coding for red are synchronized with neurons coding for motion to the left. Such content-specific synchronization between neurons has for example been observed in working-memory tasks in monkeys (Salazar et al. 2012). **(c)** Even when activity is synchronized within the color or motion area, respectively, a coherent conscious percept does not arise unless the areas are synchronized with each other.

can be formed. One attractive neural mechanism that can account for this requirement is neuronal synchrony (Bosman et al. 2012; Bressler et al. 1993; Salazar 2012). As has been discussed in greater detail elsewhere (Melloni & Singer 2010; Melloni this collection; Singer this collection), areas can be brought into *direct* contact with each other by synchronizing their neuronal activity, for example by phase alignment of neuronal oscillations, thus binding them into a functionally coherent assembly that forms a distributed representation of perceptual content. This self-organizing process can flexibly create and dissolve assemblies on top of a fixed anatomical architecture and does so without the need for anatomical convergence or broadcasting bottlenecks.

For example, imagine that a subject is confronted with two superimposed, transparent surfaces of moving dots (Figure 1). The dots on one surface are green and move to the left, and the dots on the other surface are red and move to the right. The two colors of dots are represented in a brain area coding for color, while the two motion directions are represented in an area coding for motion. For the subject to become conscious of the two surfaces, the neurons coding for green in the color area would need to synchronize their activity with the neurons coding for motion to the left in the motion area, and the neurons coding for red would need to synchronize with the neurons coding for motion to the right (Figure 1a). If the dots change direction, a new state of synchronization needs to be established, this time linking neurons coding for green with neurons coding for motion to the right, and neurons coding for red with neurons coding for motion to the left (Figure 1b). Hence, while the contents of the subject's experience are determined by the specific neuronal assemblies being active, conscious perception would be an emergent property of the state of synchronization. Recent tracing and modelling work in the macaque brain suggests that the kind of direct connectivity required to flexibly instantiate numerous, high-dimensional combinations of features is indeed afforded by high-density, reciprocal connections between brain areas (Markov et al. 2013).

Theoretical considerations and empirical evidence further suggest that the critical feature

differentiating conscious from unconscious processing is the spatial scale at which information is exchanged: while the integration of information in local modules, even in higher sensory areas (Sterzer et al. 2008), does not give rise to conscious experience by itself, large-scale integration over long distances does (Del Cul et al. 2007; Melloni et al. 2007). In the example of the transparent surfaces, this implies that even when activity is synchronized within the color or motion area, respectively, a coherent conscious percept cannot arise unless the areas are synchronized with each other (Figure 1c). Taken together, functional connectivity between distributed brain areas (i.e., connectivity that does not imply that one drives or controls the other) is an attractive alternative to localization in PFC as a candidate for the neural correlate of consciousness.

Coming back to the MVPA technique, this proposal makes a clear prediction that could be tested using decoding algorithms: specifically, one would predict that the large-scale connectivity patterns *between* brain regions for different percepts should differ, even if only slightly, for different conscious contents, and hence that conscious content should be decodable from them. This may well be the case in light of the fact that, at a much coarser scale, the neural correlates of auditory and visual awareness involve different brain networks (Eriksson et al. 2007), and that higher decoding accuracy for a subject's percept can be achieved if the joint activity patterns of several areas are considered instead of only singular patterns (Pessoa & Padmala 2007). The MVPA technique could in principle also be applied to other neuroimaging techniques that afford higher time resolution, such as electro- or magnetoencephalography or electrocorticography, thus potentially resolving the problem that arises because the representational carriers of perceptual content are highly dynamic and thus require a time-resolved analysis.

5 Conclusions

In summary, a complete theory for the neural correlates of consciousness should be able to account for the neural implementation of the contents of consciousness. John-Dylan Haynes has proposed a

clever experimental approach to localizing the contents of consciousness in the human brain, and has found that the PFC does not seem to be involved in this representation. Although surprising at first sight, this null result lines up well with the overall inconclusive evidence for a direct involvement of the PFC in conscious perception. However, it remains to be seen whether localization is the most fruitful approach to identifying the neural correlates of consciousness, or whether a more dynamic view that embraces the importance of communication between brain areas will bring us closer to solving the enigma of consciousness in the brain.

Acknowledgements

CMS is supported by a Human Frontier Science Program Long-term Fellowship (LT001118/2012-L). I would like to thank Lucia Melloni for her continuous support, insightful discussions, and comments on this manuscript.

References

- Anderson, M. L. & Oates, T. (2010). *A critique of multi-voxel pattern analysis*. Portland: Paper presented at the 32nd Annual Meeting of the Cognitive Science Society.
- Bichot, N. P., Schall, J. D. & Thompson, K. G. (1996). Visual feature selectivity in frontal eye fields induced by experience in mature macaques. *Nature*, *381* (6584), 697-699. [10.1038/381697a0](https://doi.org/10.1038/381697a0)
- Bosman, C. A., Schoffelen, J. M., Brunet, N., Oostenveld, R., Bastos, A. M., Womelsdorf, T., Rubehn, B., Stieglitz, T., De Weerd, P. & Fries, P. (2012). Attentional stimulus selection through selective synchronization between monkey visual areas. *Neuron*, *75* (5), 875-888. [10.1016/j.neuron.2012.06.037](https://doi.org/10.1016/j.neuron.2012.06.037)
- Botvinick, M. M. (2008). Hierarchical models of behavior and prefrontal function. *Trends in Cognitive Sciences*, *12* (5), 201-208. [10.1016/j.tics.2008.02.009](https://doi.org/10.1016/j.tics.2008.02.009)
- Braun, A. R., Balkin, T. J., Wesensten, N. J., Gwadry, F., Carson, R. E., Varga, M., Baldwin, P., Belenky, G. & Herscovitch, P. (1998). Dissociated pattern of activity in visual cortices and their projections during human rapid eye movement sleep. *Science*, *279* (5347), 91-95. [10.1126/science.279.5347.91](https://doi.org/10.1126/science.279.5347.91)
- Bressler, S. L., Coppola, R. & Nakamura, R. (1993). Episodic multiregional cortical coherence at multiple frequencies during visual task performance. *Nature*, *366* (6451), 153-156. [10.1038/366153a0](https://doi.org/10.1038/366153a0)
- Dehaene, S. & Changeux, J. P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, *70* (2), 200-227. [10.1016/j.neuron.2011.03.018](https://doi.org/10.1016/j.neuron.2011.03.018)
- Dehaene, S., Naccache, L., Cohen, L., Bihan, D. L., Mangin, J. F., Poline, J. B. & Riviere, D. (2001). Cerebral mechanisms of word masking and unconscious repetition priming. *Nature Neuroscience*, *4* (7), 752-758. [10.1038/89551](https://doi.org/10.1038/89551)
- Del Cul, A., Baillet, S. & Dehaene, S. (2007). Brain dynamics underlying the nonlinear threshold for access to consciousness. *PLoS Biology*, *5* (10), e260. [10.1371/journal.pbio.0050260](https://doi.org/10.1371/journal.pbio.0050260)
- Del Cul, A., Dehaene, S., Reyes, P., Bravo, E. & Slachevsky, A. (2009). Causal role of prefrontal cortex in the threshold for access to consciousness. *Brain*, *132* (9), 2531-2540. [10.1093/brain/awp111](https://doi.org/10.1093/brain/awp111)
- Desseilles, M., Dang-Vu, T. T., Sterpenich, V. & Schwartz, S. (2011). Cognitive and emotional processes during dreaming: A neuroimaging view. *Consciousness and Cognition*, *20* (4), 998-1008. [10.1016/j.concog.2010.10.005](https://doi.org/10.1016/j.concog.2010.10.005)
- Diaz, M. T. & McCarthy, G. (2007). Unconscious word processing engages a distributed network of brain regions. *Journal of Cognitive Neuroscience*, *19* (11), 1768-1775. [10.1162/jocn.2007.19.11.1768](https://doi.org/10.1162/jocn.2007.19.11.1768)
- Eriksson, J., Larsson, A., Ahlstrom, K. R. & Nyberg, L. (2007). Similar frontal and distinct posterior cortical regions mediate visual and auditory perceptual awareness. *Cerebral Cortex*, *17* (4), 760-765. [10.1093/cercor/bhk029](https://doi.org/10.1093/cercor/bhk029)
- Freeman, J., Brouwer, G. J., Heeger, D. J. & Merriam, E. P. (2011). Orientation decoding depends on maps, not columns. *Journal of Neuroscience*, *31* (12), 4792-4804. [10.1523/JNEUROSCI.5160-10.2011](https://doi.org/10.1523/JNEUROSCI.5160-10.2011)
- Funahashi, S., Bruce, C. J. & Goldman-Rakic, P. S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology*, *61* (2), 331-349.
- Grosbras, M. H. & Paus, T. (2003). Transcranial magnetic stimulation of the human frontal eye field facilitates visual awareness. *European Journal of Neuroscience*, *18* (11), 3121-3126. [10.1111/j.1460-9568.2003.03055.x](https://doi.org/10.1111/j.1460-9568.2003.03055.x)
- Hagler, D. J. jr. & Sereno, M. I. (2006). Spatial maps in frontal and prefrontal cortex. *NeuroImage*, *29* (2), 567-577. [10.1016/j.neuroimage.2005.08.058](https://doi.org/10.1016/j.neuroimage.2005.08.058)

- Haynes, J. D. (2009). Decoding visual consciousness from human brain signals. *Trends in Cognitive Sciences*, 13 (5), 194-202. [10.1016/j.tics.2009.02.004](https://doi.org/10.1016/j.tics.2009.02.004)
- (2015). An information-based approach to consciousness: Mental state decoding. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Heath, R. G., Carpenter, M. B., Mettler, F. A. & Kline, N. S. (1949). Visual apparatus: Visual fields and acuity, color vision, autokinesis. *Selective partial ablation of the frontal cortex, a correlative study of its effects on human psychotic subjects* (pp. 489-491). New York, NY: Paul B. Hoeber.
- Jackendoff, R. S. (1987). *Consciousness and the computational mind*. Cambridge, MA: MIT Press.
- Jacobs, B., Schall, M., Prather, M., Kapler, E., Driscoll, L., Baca, S., Jacobs, J., Ford, K., Wainwright, M. & Trembl, M. (2001). Regional dendritic and spine variation in human cerebral cortex: A quantitative golgi study. *Cerebral Cortex*, 11 (6), 558-571. [10.1093/cercor/11.6.558](https://doi.org/10.1093/cercor/11.6.558)
- Kamback, M. C. (1973). Detection of brief light flashes by monkeys (*Macaca nemestrina*) with dorsolateral frontal ablations. *Neuropsychologia*, 11 (3), 325-329. [10.1016/0028-3932\(73\)90044-4](https://doi.org/10.1016/0028-3932(73)90044-4)
- Kamitani, Y. & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, 8 (5), 679-685. [10.1038/nn1444](https://doi.org/10.1038/nn1444)
- Latto, R. & Cowey, A. (1971). Visual field defects after frontal eye-field lesions in monkeys. *Brain Research*, 30 (1), 1-24. [10.1016/0006-8993\(71\)90002-3](https://doi.org/10.1016/0006-8993(71)90002-3)
- Lau, H. C. & Passingham, R. E. (2006). Relative blindness in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Sciences*, 103 (49), 18763-18768. [10.1073/pnas.0607716103](https://doi.org/10.1073/pnas.0607716103)
- (2007). Unconscious activation of the cognitive control system in the human prefrontal cortex. *Journal of Neuroscience*, 27 (21), 5805-5811. [10.1523/JNEUROSCI.4335-06.2007](https://doi.org/10.1523/JNEUROSCI.4335-06.2007)
- Lau, H. C. & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*, 15 (8), 365-373. [10.1016/j.tics.2011.05.009](https://doi.org/10.1016/j.tics.2011.05.009)
- Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503 (7474), 78-84. [10.1038/nature12742](https://doi.org/10.1038/nature12742)
- Markov, N. T., Ercsey-Ravasz, M., Van Essen, D. C., Knoblauch, K., Toroczkai, Z. & Kennedy, H. (2013). Cortical high-density counterstream architectures. *Science*, 342 (6158), 1238406-1238406. [10.1126/science.1238406](https://doi.org/10.1126/science.1238406)
- Markowitsch, H. J. (1995). Cerebral bases of consciousness: A historical view. *Neuropsychologia*, 33 (9), 1181-1192. [10.1016/0028-3932\(95\)00057-A](https://doi.org/10.1016/0028-3932(95)00057-A)
- Markowitsch, H. J. & Kessler, J. (2000). Massive impairment in executive functions with partial preservation of other cognitive functions: The case of a young patient with severe degeneration of the prefrontal cortex. *Experimental Brain Research*, 133 (1), 94-102. [10.1007/s002210000404](https://doi.org/10.1007/s002210000404)
- Melloni, L. (2015). Consciousness as inference in time—A commentary on Victor Lamme. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Melloni, L., Molina, C., Pena, M., Torres, D., Singer, W. & Rodriguez, E. (2007). Synchronization of neural activity across cortical areas correlates with conscious perception. *Journal of Neuroscience*, 27 (11), 2858-2865. [10.1523/JNEUROSCI.4623-06.2007](https://doi.org/10.1523/JNEUROSCI.4623-06.2007)
- Melloni, L. & Singer, W. (2010). Distinct characteristics of conscious experience are met by large scale neuronal synchronization. In E. Perry, D. Collerton, F. E. N. Le-Beau & H. Ashton (Eds.) *New horizons in the neuroscience of consciousness* (pp. 17-28). Amsterdam, NL: John Benjamins.
- Muzur, A., Pace-Schott, E. F. & Hobson, J. A. (2002). The prefrontal cortex in sleep. *Trends in Cognitive Science*, 6 (11), 475-481. [10.1016/s1364-6613\(02\)01992-7](https://doi.org/10.1016/s1364-6613(02)01992-7)
- Nakamura, R. K. & Mishkin, M. (1986). Chronic 'blindness' following lesions of nonvisual cortex in the monkey. *Experimental Brain Research*, 63 (1), 173-184. [10.1007/BF00235661](https://doi.org/10.1007/BF00235661)
- Panagiotaropoulos, T. I., Deco, G., Kapoor, V. & Logothetis, N. K. (2012). Neuronal discharges and gamma oscillations explicitly reflect visual consciousness in the lateral prefrontal cortex. *Neuron*, 74 (5), 924-935. [10.1016/j.neuron.2012.04.013](https://doi.org/10.1016/j.neuron.2012.04.013)
- Passingham, R. (2009). How good is the macaque monkey model of the human brain? *Current Opinion in Neurobiology*, 19 (1), 6-11. [10.1016/j.conb.2009.01.002](https://doi.org/10.1016/j.conb.2009.01.002)
- Pessoa, L. & Padmala, S. (2007). Decoding near-threshold perception of fear from distributed single-trial brain activation. *Cerebral Cortex*, 17 (3), 691-701. [10.1093/cercor/bhk020](https://doi.org/10.1093/cercor/bhk020)
- Rainer, G., Asaad, W. F. & Miller, E. K. (1998). Memory fields of neurons in the primate prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 95 (25), 15008-15013. [10.1073/pnas.95.25.15008](https://doi.org/10.1073/pnas.95.25.15008)

- Rigotti, M., Barak, O., Warden, M. R., Wang, X. J., Daw, N. D., Miller, E. K. & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497 (7451), 585-590. [10.1038/nature12160](https://doi.org/10.1038/nature12160)
- Roe, A. W. (2010). Optical imaging of short-term working memory in prefrontal cortex of the macaque monkey. In A. W. Roe (Ed.) *Imaging the brain with optical methods* (pp. 119-133). New York, NY: Springer.
- Roelfsema, P. R. & Treue, S. (2014). Basic neuroscience research with nonhuman primates: a small but indispensable component of biomedical research. *Neuron*, 82 (6), 1200-1204. [10.1016/j.neuron.2014.06.003](https://doi.org/10.1016/j.neuron.2014.06.003)
- Romo, R., Brody, C. D., Hernandez, A. & Lemus, L. (1999). Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature*, 399 (6735), 470-473. [10.1038/20939](https://doi.org/10.1038/20939)
- Rounis, E., Maniscalco, B., Rothwell, J. C., Passingham, R. E. & Lau, H. (2010). Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs meta-cognitive visual awareness. *Cognitive Neuroscience*, 1 (3), 165-175. [10.1080/17588921003632529](https://doi.org/10.1080/17588921003632529)
- Sahraie, A., Weiskrantz, L., Barbur, J. L., Simmons, A., Williams, S. C. & Brammer, M. J. (1997). Pattern of neuronal activity associated with conscious and unconscious processing of visual signals. *Proceedings of the National Academy of Sciences of the United States of America*, 94 (17), 9406-9411. [10.1073/pnas.94.17.9406](https://doi.org/10.1073/pnas.94.17.9406)
- Salazar, R. F., Dotson, N. M., Bressler, S. L. & Gray, C. M. (2012). Content-specific fronto-parietal synchronization during visual working memory. *Science*, 338 (6110), 1097-1100. [10.1126/science.1224000](https://doi.org/10.1126/science.1224000)
- Schwiedrzik, C. M., Ruff, C. C., Lazar, A., Leitner, F. C., Singer, W. & Melloni, L. (2014). Untangling perceptual memory: hysteresis and adaptation map into separate cortical networks. *Cerebral Cortex*, 24 (5), 1152-1164. [10.1093/cercor/bhs396](https://doi.org/10.1093/cercor/bhs396)
- Singer, W. (2015). The ongoing search for the neuronal correlate of consciousness. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Sreenivasan, K. K., Curtis, C. E. & D'Esposito, M. (2014). Revisiting the role of persistent neural activity during working memory. *Trends in Cognitive Science*, 18 (2), 82-89. [10.1016/j.tics.2013.12.001](https://doi.org/10.1016/j.tics.2013.12.001)
- Sterzer, P., Haynes, J. D. & Rees, G. (2008). Fine-scale activity patterns in high-level visual areas encode the category of invisible objects. *Journal of Vision*, 8 (15), 1-12. [10.1167/8.15.10](https://doi.org/10.1167/8.15.10)
- Stokes, M. G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D. & Duncan, J. (2013). Dynamic coding for cognitive control in prefrontal cortex. *Neuron*, 78 (2), 364-375. [10.1016/j.neuron.2013.01.039](https://doi.org/10.1016/j.neuron.2013.01.039)
- Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, 5 (42). [10.1186/1471-2202-5-42](https://doi.org/10.1186/1471-2202-5-42)
- van Gaal, S., Ridderinkhof, K. R., Fahrenfort, J. J., Scholte, H. S. & Lamme, V. A. (2008). Frontal cortex mediates unconsciously triggered inhibitory control. *The Journal of Neuroscience*, 28 (32), 8053-8062. [10.1523/JNEUROSCI.1278-08.2008](https://doi.org/10.1523/JNEUROSCI.1278-08.2008)
- Veselis, R. A., Feshchenko, V. A., Reinsel, R. A., Dnistrian, A. M., Beattie, B. & Akhurst, T. J. (2004). Thiopental and propofol affect different regions of the brain at similar pharmacologic effects. *Anesthesia & Analgesia*, 99 (2), 399-408. [10.1213/01.ANE.0000131971.92180.DF](https://doi.org/10.1213/01.ANE.0000131971.92180.DF)
- Vilarroya, O. (2013). The challenges of neural mind-reading paradigms. *Frontiers in Human Neuroscience*, 7 (306). [10.3389/fnhum.2013.00306](https://doi.org/10.3389/fnhum.2013.00306)
- von der Malsburg, C. (1997). The coherence definition of consciousness. In M. Ito, Y. Miyashita & E. T. Rolls (Eds.) *Cognition, computation and consciousness* (pp. 193-204). Oxford, UK: Oxford University Press.
- Zaksas, D. & Pasternak, T. (2006). Directional signals in the prefrontal cortex and in area MT during a working memory for visual motion task. *The Journal of Neuroscience*, 26 (45), 11726-11742. [10.1523/JNEUROSCI.3420-06.2006](https://doi.org/10.1523/JNEUROSCI.3420-06.2006)

Can Synchronization Explain Representational Content?

A Reply to Caspar M. Schwiedrzik

John-Dylan Haynes

Multivariate decoding provides an important tool for studying the representation and transformation of mental contents in the human brain. Specifically, decoding can be used to identify the neural correlates of contents of consciousness (NCCs). Decoding of functional magnetic resonance imaging (fMRI) signals has so far mostly revealed content-selectivity in sensory brain regions, but not in prefrontal cortex. The limitations of fMRI-decoding only permit cautious conclusions because fMRI signals are only indirectly related to neural coding. However, the role of prefrontal cortex in visual awareness is also questioned by other findings, reviewed in [Schwiedrzik \(this collection\)](#). Neural synchronization might offer an alternative to solving the binding problem by providing a computational means of integrating information encoded in distributed brain regions. However, it is unclear whether synchronization in itself serves as a coding dimension for visual features. Furthermore, other alternatives to synchronization, especially the role of spatial codes, need to be considered as potential solutions to the feature binding problem.

Keywords

Bias | Binding problem | Contents of consciousness | Dynamically changing coding space | Fmri-decoding | Functional magnetic resonance imaging | Global workspace theory | Multivariate decoding | Prefrontal cortex | Spatial code | Synchronization | Tolerance | Visual awareness

Author

[John-Dylan Haynes](#)

haynes@bccn-berlin.de

Charité – Universitätsmedizin Berlin
Berlin, Germany

Commentator

[Caspar M. Schwiedrzik](#)

cschwiedrz@rockefeller.edu

The Rockefeller University
New York, NY, U.S.A.

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

Information-based approaches to brain function have been very successful in recent years ([Pouget et al. 2000](#); [Haynes & Rees 2006](#); [Kriegeskorte et al. 2006](#)). Most importantly, they allow to study how mental contents are represented and transformed during information processing in the brain. My target article in this volume ([Haynes this collection](#)) emphasized the importance of an information-based approach for the study of human consciousness, especially for understanding the neural mechanisms of

visual awareness. Whereas many previous studies mainly aimed to establish which additional processing needs to occur for a stimulus to reach awareness, a second question is equally important: how and where the brain encodes the specific contents of consciousness. Research on these neural correlates of the contents of consciousness (NCCs; [Chalmers 2000](#); [Block 2007](#); [Koch 2004](#)) has been sparse. For identifying NCCs, simply establishing that a brain area responds stronger under conscious than un-

der unconscious processing is not sufficient, because this could merely reflect unspecific processes such as attention or memory (Corbetta & Shulman 2002; Goldman-Rakic 1995). Instead, for identifying the neural code of contents several specific questions need to be addressed: Which brain regions encode sensory information in a representational space that exactly matches perception? And under which circumstances does a crossing of the threshold to awareness involve changes of representations in these specialized coding spaces?

2 The role of prefrontal cortex

One example of the importance of considering content-based processing is the global workspace theory (Dehaene & Naccache 2001; Baars 2002). In specific readings of the theory, consciousness involves a distribution of sensory information from sensory cortices to parietal and prefrontal cortex (Haynes this collection). Increased activity in frontoparietal regions under conscious perception is seen as evidence for such a “broadcasting” of sensory information (Baars 2002). However, without additional support by information-based or representational analyses, increased activity in frontoparietal regions with increased awareness might simply reflect unspecific processes, say as in detecting or reporting a change in perception, rather than coding the sensory information itself. In several studies with functional magnetic resonance imaging (fMRI) we found no evidence for changes in prefrontal representation of sensory information under increased levels of awareness (reviewed in Haynes this collection). Thus, we found no evidence that sensory information is re-represented in prefrontal or parietal cortex.

At this important point, the comment by Schwiedrzik (this collection) adds further important details on the potential role of prefrontal cortex (PFC) in visual awareness (Crick & Koch 1995; Dehaene & Naccache 2001). In a first line of arguments Schwiedrzik provides more detail on a point briefly sketched in the original article (Haynes this collection), whether absence of decodable information in PFC might reflect limitations of fMRI-based pattern classi-

fication. fMRI decoding will only be able to access neural information that is encoded in specific formats and topologies (Chaimow et al. 2011). For example, if neurons with different tuning properties are randomly distributed within a voxel, then the voxel will not be able to pick up any information about these properties. Thus, a macroscopic clustering of cells with similar tuning properties is required for fMRI-decoding to pick up information. So information could be present in prefrontal cortex, but in a format that is not accessible to fMRI, not even with the increased sensitivity of multivariate analyses. This might explain the discrepancy between the absence of information in many fMRI studies and differential responsivity to stimulus features of cells in PFC in non-human primates (Pasternak & Greenlee 2005). Please note, however, that most of the evidence for sensory tuning in PFC is obtained under working memory paradigms, which also includes temporal two-alternative forced choice tasks (Romo et al. 1999). Thus, it is unclear whether this generalizes to “realtime” perceptual experience.

There are further challenges in accessing neural information. Schwiedrzik (this collection) brings forward an important point already raised previously (Duncan & Owen 2000): coding in prefrontal cortex might be dynamic and thus the code might change across time. Such dynamically changing coding spaces might again not be detectable in classification analyses that assume a constant population code across the period analysed (Stokes et al. 2013). These points raised by Schwiedrzik are fully valid: It is highly important to consider these limitations when interpreting the results of fMRI decoding studies. To some degree these challenges might be alleviated with future technical developments. For example, columnar-level information can be accessed following recent advances in high-resolution fMRI (Yacoub et al. 2008). However, many limitations of fMRI will remain due to its vascular origin that only samples neural information indirectly. Please note that the limitations go far beyond the points raised by Schwiedrzik (this collection). For example, fMRI might not only miss information, but it might

also tap into information that is not available at the level of single neurons. Say, if an fMRI voxel samples a homogenous group of cells with highly similar tuning properties, the voxel might reflect a degree of averaging that is not available at the level of single neurons.

Please note, that the target article was not restricted to decoding approaches in fMRI alone. Instead, the aim was to outline a more general approach to studying the neural correlates of the contents of consciousness. If suitable recording techniques were available the information could be assessed based on a whole family of potential representational signals, including especially axonal and dendritic population activity. Please further note, that any recording technique has its blind spots. For example single-cell electrophysiology is biased towards large cells (Bartels et al. 2008), or optical imaging with voltage sensitive dyes is restricted to the surface of the brain (Grinvald & Hildesheim 2004). Thus, only a combination of techniques will be able to provide a full picture of the changes in neural coding with varying levels of visual awareness. Importantly, in a second line of argumentation Schwiedrzik (this collection) provides additional support and plausibility to the finding that the absence of information in PFC is real. For example the effects of lesions in PFC on visual recognition can be strikingly weak, which would not be expected if representation or routing of information in PFC were a necessary condition for awareness.

A third line of argumentation brings neural integration between spatially separated brain regions into play as a different potential mechanism of visual awareness (Engel & Singer 2001). The basic idea is that representation might involve a dynamic binding involving a synchronization of neural activity. According to this model, as Schwiedrzik points out, a distribution of sensory information into prefrontal brain regions might not be necessary. Many studies have related changes in neural synchronization to changes in visual awareness (see e.g., Engel & Singer 2001; Uhlhaas et al. 2009). However, it is important to look more closely at the explanation that can be obtained by changes in synchronization, specifically if the aim is to

provide an explanation of the neural correlates of contents of consciousness (NCCCs). Typically, synchronization is not viewed as a coding dimension for contents, but as a code for binding and integration of features that are distributed across multiple content-specific regions (von der Malsburg 1999). The example provided in figure 1 of the comment by Schwiedrzik illustrates this nicely. A person views two superimposed clouds of moving dots, one green cloud moving left and a red cloud moving right. The features are encoded in content-specific fashion with two different activation patterns in the color area coding the two colors and two different activation patterns in the motion area coding the two motion directions. Synchronization between the neural representations of “green” and “left” on the one hand and “red” and “right” enables a separate binding of these two distributed features and also allows them to jointly be more effective in activating any downstream brain regions (König et al. 1996). Here, the contents are represented as differential activation states in the content-specific regions and their binding is achieved by synchronization. In this example representation and binding are separable problems based on separate computational mechanisms. However, Schwiedrzik (this collection) also goes one step further by suggesting that the large-scale connectivity patterns between brain regions might themselves code for different conscious contents. It has already been shown with fMRI that connectivity patterns between remote brain regions reflect changes in visual awareness (e.g., Haynes et al. 2005; Imamoglu et al. 2012). It has also been shown that connectivity matrices obtained with fMRI can be used to classify cognitive states (Richiardi et al. 2011; Heinzle et al. 2012). However, to date I am not aware of any evidence that fine-grained perceptual contents are encoded in differential patterns of brain connectivity. Furthermore, synchronization sometimes fails to explain perception (Thiele & Stoner 2003) and there other solutions to the binding problem besides synchronization. For example, high-level content-selective brain regions that achieve a certain degree of tolerance to variations in spatial location, still have con-

siderable information about the spatial location of features (e.g., [Cichy et al. 2011](#)). Thus, the spatial maps and their associated differential anatomical (as opposed to functional) connectivity patterns provide a plausible alternative hypothesis to synchronization ([Treisman & Gelade 1980](#)).

3 Conclusions

Prefrontal cortex is often considered vital for visual awareness ([Crick & Koch 1995](#); [Dehaene & Naccache 2001](#)), however multivariate decoding studies have revealed a marked absence of sensory information in prefrontal cortex ([Haynes this collection](#)). Neural synchronization ([Schwiedrzik this collection](#)) might provide an alternative account for feature binding and selective routing of information. However, it is currently unclear whether any form of functional connectivity can itself code specific sensory contents.

References

- Baars, B. J. (2002). The conscious access hypothesis: Origins and recent evidence. *Trends in Cognitive Sciences*, 6 (1), 47-52.
- Bartels, A., Logothetis, N. K. & Moutoussis, K. (2008). fMRI and its interpretations: An illustration on directional selectivity in area V5/MT. *Trends in Neurosciences*, 31 (9), 444-453. [10.1016/j.tins.2008.06.004](https://doi.org/10.1016/j.tins.2008.06.004)
- Block, N. (2007). Consciousness, accessibility and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences*, 30 (5-6), 481-499. [10.1017/S0140525X07002786](https://doi.org/10.1017/S0140525X07002786)
- Chaimow, D., Yacoub, E., Ugurbil, K. & Shmuel, A. (2011). Modeling and analysis of mechanisms underlying fMRI-based decoding of information conveyed in cortical columns. *NeuroImage*, 56 (2), 627-642. [10.1016/j.neuroimage.2010.09.037](https://doi.org/10.1016/j.neuroimage.2010.09.037)
- Chalmers, D. (2000). What is a neural correlate of consciousness? In T. Metzinger (Ed.) *Neural correlates of consciousness: Conceptual and empirical questions* (pp. 17-40). Boston, MA: MIT.
- Cichy, R. M., Chen, Y. & Haynes, J. D. (2011). Encoding the identity and location of objects in human LOC. *NeuroImage*, 54 (3), 2297-2307. [10.1016/j.neuroimage.2010.09.044](https://doi.org/10.1016/j.neuroimage.2010.09.044)
- Corbetta, M. & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3 (3), 201-215.
- Crick, F. & Koch, C. (1995). Are we aware of neural activity in primary visual cortex? *Nature*, 375 (6527), 121-123.
- Dehaene, S. & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, 79 (1-2), 1-37.
- Duncan, J. & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences*, 23 (10), 475-483.
- Engel, A. K. & Singer, W. (2001). Temporal binding and the neural correlates of sensory awareness. *Trends in Cognitive Sciences*, 5 (1), 16-25.
- Goldman-Rakic, P. S. (1995). Cellular basis of working memory. *Neuron*, 14 (3), 477-485.
- Grinvald, A. & Hildesheim, R. (2004). VSDI: A new era in functional imaging of cortical dynamics. *Nature Reviews Neuroscience*, 5 (11), 874-885.
- Haynes, J. D. (2015). An information-based approach to consciousness: Mental state decoding. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a.M., GER: MIND Group.

- Haynes, J. D., Driver, J. & Rees, G. (2005). Visibility reflects dynamic changes of effective connectivity between V1 and fusiform cortex. *Neuron*, 46 (5), 811-821.
- Haynes, J. D. & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7 (7), 523-534.
- Heinze, J., Wenzel, M. A. & Haynes, J. D. (2012). Visuo-motor functional network topology predicts upcoming tasks. *Journal of Neuroscience*, 9960 (9968), 475-483. [10.1523/JNEUROSCI.1604-12.2012](https://doi.org/10.1523/JNEUROSCI.1604-12.2012)
- Imamoglu, F., Kahnt, T., Koch, C. & Haynes, J. D. (2012). Changes in functional connectivity support conscious object recognition. *NeuroImage*, 63 (4), 1909-1917. [10.1016/j.neuroimage.2012.07.056](https://doi.org/10.1016/j.neuroimage.2012.07.056)
- Koch, C. (2004). *The quest for consciousness: A neurobiological approach*. Englewood: Roberts.
- Kriegeskorte, N., Goebel, R. & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the USA*, 103 (10), 3863-3868.
- König, P., Engel, A. K. & Singer, W. (1996). Integrator or coincidence detector? The role of the cortical neuron revisited. *Trends in Neurosciences*, 19 (4), 130-137.
- Pasternak, T. & Greenlee, M. W. (2005). Working memory in primate sensory systems. *Nature Reviews Neuroscience*, 6 (2), 97-107.
- Pouget, A., Dayan, P. & Zemel, R. (2000). Information processing with population codes. *Nature Reviews Neuroscience*, 1 (2), 125-132.
- Richiardi, J., Eryilmaz, H., Schwartz, S., Vuilleumier, P. & Van De Ville, D. (2011). Decoding brain states from fMRI connectivity graphs. *NeuroImage*, 56 (2), 616-626. [10.1016/j.neuroimage.2010.05.081](https://doi.org/10.1016/j.neuroimage.2010.05.081)
- Romo, R., Brody, C. D., Hernández, A. & Lemus, L. (1999). Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature*, 399 (6735), 470-473.
- Schwiedrzik, C. (2015). What's up with prefrontal cortex? – A commentary on John-Dylan Haynes. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a.M., GER: MIND Group.
- Stokes, M. G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D. & Duncan, J. (2013). Dynamic coding for cognitive control in prefrontal cortex. *Neuron*, 78 (2), 364-375.
- Thiele, A. & Stoner, G. (2003). Neuronal synchrony does not correlate with motion coherence in cortical area MT. *Nature*, 421 (6921), 366-370.
- Treisman, A. M. & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12 (1), 97-136.
- Uhlhaas, P. J., Pipa, G., Lima, B., Melloni, L., Neuenschwander, S., Nikolic, D. & Singer, W. (2009). Neural synchrony in cortical networks: History, concept and current status. *Frontiers in Integrative Neuroscience*, 3 (17). [10.3389/neuro.07.017.2009](https://doi.org/10.3389/neuro.07.017.2009)
- von der Malsburg, C. (1999). The what and why of binding: The modeler's perspective. *Neuron*, 24 (1), 95-104.
- Yacoub, E., Harel, N. & Ugurbil, K. (2008). High-field fMRI unveils orientation columns in humans. *Proceedings of the National Academy of Sciences of the USA*, 105 (30), 10607-10612. [10.1073/pnas.0804110105](https://doi.org/10.1073/pnas.0804110105)

Beyond Illusions

On the Limitations of Perceiving Relational Properties

Heiko Hecht

Explaining the perception of our visual world is a hard problem because the visual system has to fill the gap between the information available to the eye and the much richer visual world that is derived from the former. Perceptual illusions continue to fascinate many researchers because they seem to promise a glimpse of how the visual system fills this gap. Illusions are often interpreted as evidence of the error-prone nature of the process. Here I will show that the opposite is true. To do so, I introduce a novel stance on what constitutes an illusion, arguing that the traditional view (illusion as mere discrepancy between stimulus and percept) has to be replaced by illusion as a manifest noticed discrepancy. The two views, unfortunately, are not necessarily related. On the contrary; we experience the most spectacular illusions where our perception is pretty much on target. Once our interpretation of the sensory data is off the mark, we usually no longer experience illusions but live happily without ever noticing the enormous perceptual and conceptual errors we make. The farther we move away from simple pictorial stimuli as the subject of our investigations, the more commonplace a discrepancy between percept and reality does become—and the less likely we are willing to call it illusory. Two case studies of our perception of relational properties will serve to illustrate this idea. The case studies are based on the conviction that perceiving is more than mere sensation, and that some degree of (unconscious) judgment is a necessary ingredient of perception. We understand little about how to balance objects and we make fundamental mistakes when perceiving the slipperiness of surfaces. All the while, we never experience illusions in this context. Thus, when dealing with simple percepts, illusions may be revealing. But when it comes to percepts that involve relational properties, illusions fail to arise, as perception is not concerned with veridicality but appears to be satisfied with the first solution that does not interfere with our daily activities.

Keywords

Error | Illusion | Intuitive physics | Underspecification

Author

Heiko Hecht

hecht@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Commentator

Axel Kohler

axelkohler@web.de

Universität Osnabrück
Osnabrück, Germany

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Illusion?

1.1 The underspecification problem (UP)

Visual perception can be seen as the process by which the visual system interprets the sensory core data that come in through the retinæ of the eyes (see e.g., [Hatfield & Epstein 1979](#)). The sensory core is not sufficient to specify the percept; that is, there is an explanatory gap between the information present at the retina—which is in essence two-dimensional (2D)—and the information present in the three-dimensional (3D) objects that we see. Let us call the prob-

lem that arises in having to fill this gap the “underspecification problem” (see [Hecht 2000](#)). Figure 1 illustrates the UP (underspecification problem). A given object can only project one particular image onto the projection surface (retina); however, a given projection could have been caused by an indefinite number of objects in the world. Because of this anisotropy in the mapping between the 3D object and its 2D projection, information is lost during the projective process, which cannot be regained with certainty. One could argue that the history of per-

ception theories is more or less the history of finding solutions to reconstruct the 3D object that has caused a given projection.

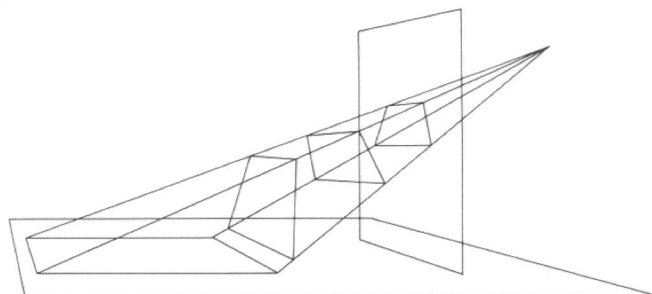


Figure 1: Underspecification: The 3D origin of a given image on the retina (here approximated by the vertical projection screen) is provided by an indefinite number of objects at various orientations in space. Illustration from [Gibson \(1979\)](#).

In order to assess the quality of the solution offered by a given perceptual theory, we have to evaluate how it describes the gap between sensory core and percept and the mechanism by which it suggests that the gap is being bridged. The Gibsonian theory of direct perception aside—which denies the problem altogether (e.g., [Gibson 1979](#))—we have a variety of theories to choose from. They are all constructionist in the sense that the sensory data have to be interpreted and arranged into the configuration that is most likely or most logical. The theories differ in the mechanisms they make responsible for the reconstructive process. For instance, [Hermann von Helmholtz \(1894\)](#) supposes inferences of unconscious nature that arrive inductively or maybe abductively at a preferred solution. [Roger Shepard \(1994\)](#), on the other hand supposes a recurrence to phylogenetically-acquired knowledge. He takes the regularities of the physical world or of geometry to have been internalized through the course of evolution and to be used to disambiguate competing solutions. An example of such internalized knowledge is the fact that light usually comes from above (see [Figure 2](#)). A shading gradient from light (at the top of an object) to dark (at its bottom) would thus be compatible with a convex but not with a concave object.

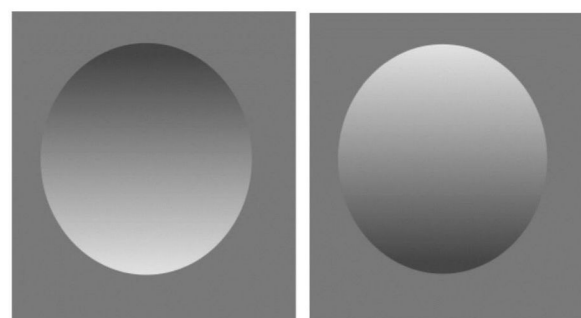


Figure 2: Solution of the underspecification by drawing on internalized knowledge that light comes from above. The sphere in the right panel looks convex because it is lighter at the top, whereas the same image rotated by 180° (left panel) looks concave. Have we created an illusion by juxtaposing them?

Others have proposed that the system considers statistical probabilities by defaulting to contextually appropriate, high-frequency responses ([Reason 1992](#)) or by applying the Bayes-theorem (e.g., [Knill & Richards 1996](#); [Kersten et al. 2004](#)), or predictive processing ([Clark this collection](#); [Hohwy this collection](#)). Here we are not concerned with the exact nature of how the construction is accomplished. Note, however, that all the solutions that have been proposed abound with cognitive ingredients. The process of constructing a 3D object from the 2D retinal input is usually thought to draw on memory and on some sort of inferencing, albeit unconsciously. The next step to arriving at meaningful percepts on the basis of the 3D object—which is just as essential in perception—involves even more cognitive elements, be they unconscious or amenable to consciousness.

Here I would like to include a brief aside, which may seem obvious to the psychologist but not so obvious to the philosopher. Perceiving cannot be dissected successfully into a sensational part and a judgmental part when we are dealing with the everyday perception of meaningful objects. Perceiving is always judgmental when we see a stick or a bird, or when it comes to seeing that we can pick up the stick and that it falls down when we release it. In other words, pure sensations may be possible introspectively—sensing red, sensing heat etc.—but they are

no longer possible in everyday object perception, that is a separation of sensation and judgment is not ecologically valid. Take, for instance, the falling object as given in phenomenal perception. In the sub-field of experimental psychology called “intuitive physics”, investigators have doctored physical events to contradict Newtonian physics and presented visual animations to novice or expert observers. Many of the latter do not see anything wrong with objects falling straight down when released, as opposed to following the proper parabola that they should (see section 1.3.1 on so-called cognitive illusions). This perception is reflected in motor action—people release the object in the wrong place when trying to hit a container; this perception arises in toddlers unable to reflect upon the event, and it persists after formal physics training in cases where observers have to make quick decisions. Thus, a separation into a sensation and perceptual judgment is not meaningful here. Perception of (everyday) objects and events necessarily includes a judgmental aspect, which may or may not enter consciousness.

Now, we are concerned with the question of whether the errors that arise during the perceptual process can be used to gauge where the visual system fails to capture the 3D world. We will argue that this is not the case. Research focusing on so-called optical illusions is particularly ill-suited to gain insight into how the visual system solves the UP. Illusions typically arise when errors are rather small, thus the presence or magnitude of an illusion is no predictor of the size of the UP. By and large, perceptual error is rather small when it comes to simple object properties, such as size, distance, direction of motion, etc. Errors become much larger, more interesting, and potentially dangerous when it comes to relational properties, such as seeing if an object can be lifted or if I will slip and fall when treading on a given surface. The case studies below will show that in the context of relational properties we make errors but we do not experience illusions.

1.2 The Luther illusion

Please take a close look at this painting of Martin Luther. You have certainly seen pictures of

the great protestant reformer before. Does anything about this painting strike you as strange?



Figure 3: Martin Luther as painted by Lukas Cranach the Elder (1529), Hessisches Landesmuseum Darmstadt.

You may have found that he looks well nourished, as is appropriate for a monk whose enjoyment of worldly pleasures is well documented. However, I am sure you did not notice the illusion. Well, I have photoshopped the photograph and made it 15% wider than it should be. There is a discrepancy between the painting (or veridical photograph thereof) and the picture presented in Figure 3. Such discrepancies are typically considered to be the essence of illusion. For instance, [Martinez-Conde & Macknik \(2010, p. 4\)](#) define an illusion as “the dissociation between the physical reality and the subjective perception of an object or event”. The physical reality of the picture is distorted by 15%, but your perception was that of a correct rendition of a famous painting. Now let us add another twist to the Luther illusion (Figure 4).



Figure 4: Martin Luther right side up and upside down.

Have I taken the original photograph or have I turned around the 15% wider version? Surely, Luther looks to be slimmer in the panel on the right. If you turn the page upside down, you will see that both panels show the same picture that is 15% wider than the original. Let us assume that the inversion effect—also named fat-face-thin-illusion by Peter Thompson (Thompson & Wilson 2012)—is exactly 15 % in magnitude. Has the illusion that I introduced initially been nullified by the inversion?

The fictitious Luther illusion is meant to make the point that the mere discrepancy between physical reality and a percept should not be conceived of as illusory. It may not even be reasonable to conceive of it as an error. The stretched image may be a better representation of what we know about Luther than the “correct” picture. For instance, the picture may typically be viewed from an inappropriate vantage point that could make the stretched version more veridical even when compared to the actual Luther, were he teleported into our time. Take Figure 5. I have stretched Luther by another 50%. Now he seems a bit distorted, but not to an extent that would prevent us from recognizing him or from enjoying the picture. There is a fundamental property that needs to be added for something to be considered an illusion. I contend that this is a dual simultaneous percept that tells us that what we see is so and not so at the same time (for a detailed defence of this position see Hecht 2013). For an illusion¹

to be called thus, it has to be manifest immediately and perceptually. Calling something an illusion is only meaningful if it refers to a discrepancy that we can see. It is not meaningful if it refers to some error that we have to infer.



Figure 5: Martin Luther stretched by another 50%.

Take for instance the often-cited stick in the water that looks bent. The static image presented in Figure 6 is not an illusion. We see a bent stick; note that its shadow is bent as well, and without recourse to our experience of refraction that occurs where two media adjoin, we would not know if the stick were actually bent or if some effect of optics had created the percept. However, the moment we move the stick up and down we see the illusion_m. We see the stick being bent and being straight at the same time. The illusion becomes manifest. That is, the discrepancy if not contradiction between the two percepts (here the straight and the bent stick) is available in our working memory, we become aware of it, often without being able to resolve which of the two discrepant percepts is closer to reality. In the case of the stick, the location of the bending at water level reveals that

fest illusion that is perceived rather than inferred with the help of physics text books). I will only refer to illusion_m as illusion, whereas I will refer to illusion_d as mere error or discrepancy. See also the related distinction between phenomenally opaque and phenomenally transparent illusions (e.g., Metzinger 2003a, 2003b). My distinction between illusion_d and illusion_m is meant to be merely perceptual.

¹ Note that I will differentiate between illusion_d (being the old notion of discrepancy between object and percept) and illusion_m (the mani-

the stick is really straight; however, in most cases the illusion_m remains unresolved, as for example in the case of the Ebbinghaus illusion.



Figure 6: Is the stick bent?

1.3 Thesis: Illusions_m are not evidence of error but rather unmasking of error

It would make no sense to call the circles in Figure 7 an illusion_m, even if a researcher could show with a large dataset that the inner circle is reproduced 2% bigger than it was on the picture. However, as soon as we allow for a direct comparison and put a ruler to the center circles in Figure 8, the illusion_m arises (see Wundt 1898; an interactive demonstration of the Ebbinghaus illusion can be found at <http://michaelbach.de/ot/cog-Ebbinghaus/index-de.html>).

Illusions_m are perceptually immediate but they appear to require some form of comparison and judgment, which supports the argument that phenomenal perception cannot be divided into a merely sensational core and a cognitive elaboration. For instance, in the case of a Necker cube or a bi-stable apparent motion quartet, the illusion_m can become manifest by a mere deliberate shift of attention.

Given the severity of the UP, we should not be fascinated by the existence of error (illusions_d), but should instead be fascinated by the fact that our perceptions are pretty much on target most of the time. It is truly amazing that among the enormous range of possible interpretations of the retinal image, we usually pick the appropriate

one. Illusions_m are rare special cases of ubiquitous small errors that become manifest because of some coincidence or another. Note that this assessment does not only apply to visual perception but also to other sensory modalities in which sensory information has to be interpreted and integrated. For instance, the cutaneous rabbit illusion_m arises when adjacent locations on the skin of our arm are stimulated in sequence. We experience one coherent motion (a rabbit moving along our arm) rather than a sequence of unrelated taps. This “inference” can be explained by probabilistic reasoning (Goldreich 2007) and may be considered the tactile analogue of apparent motion: just as we cannot perceptually distinguish a sequence of static stimuli from real motion in the movie theater. As a matter of fact, the pauses between the intermittent frames of the movie are indispensable for motion pictures to look smooth and continuous.

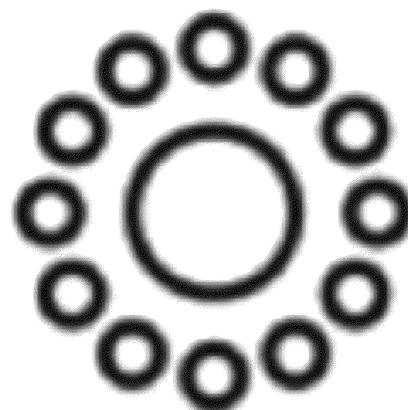


Figure 7: Is the circle in the middle perceived to be bigger than it really is? Possibly an illusion_d.

Gestalt psychologists have described the constructive process by which meaningful objects emerge from the various elements in our sensory core (see e.g., Max Wertheimer 1912 for the case of apparent motion). For good reason, they have avoided the term illusion, and introduced the term emergent property for the phenomenal result of the (unconscious) process of perceptual organization. It would violate our everyday experience to call something we see an illusion just because we know a little

bit about the underlying physics. Just because we know that our continuous motion percept is derived from a sequence of discrete images, this does not make the percept an illusion (neither illusion_m nor illusion_d).² By the same token, knowing that light is a wave (or a stream of photons) does not make objects in the world illusory. In fact, a discrepancy between what is really there and what we perceive is the norm, not the exception. Given my conceptual distinction, I will show how the perceptual system deals with the ubiquitous discrepancy, with the normal case of illusion_d . The relatively rare cases illusions_m arise as a by-product of this process. For something to deserve the name illusion, this discrepancy has to become manifest. The Ebbinghaus illusion only turns into an illusion_m when we perceive a conflict, when the inner circles are seen (or inferred) to be equal in size and they look different in size at the same time. Thus, it is not the ubiquitous presence of error that makes an illusion_m but the rather unusual case where this error is unmasked by a perceptual comparison process.

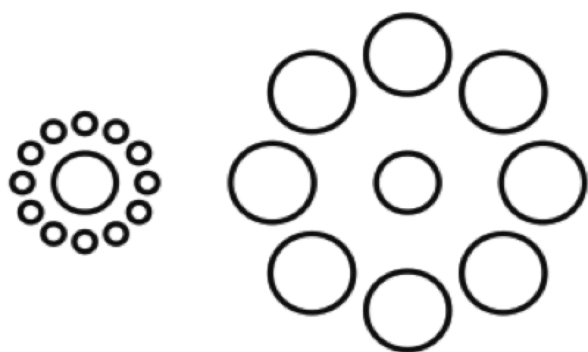


Figure 8: Is the circle surrounded by smaller circles perceived to be bigger than it really is, or is the center circle on the right perceived to be too small? This is the famous Titchener illusion_m that was invented by Hermann Ebbinghaus and first reported by Wilhelm Wundt (1898).

² Note, that a discrepancy between stimulus and percept is necessary but not sufficient for an illusion_m . Thus, all illusions require an illusion_d but will only become illusions_m in some cases. My distinction is capable of sorting out illusions as relevant to perceptual psychology, it does, however, not speak to the question of how we can describe the physical stimulus in the first place, i.e., the grand illusion argument (see <http://www.imprint.co.uk/books/noe.html>).

1.3.1 A note on so-called cognitive illusions_d

In our everyday perception, once we consider that objects are often in motion and carry meaning at the perceptual level (see Gibson's concept of affordance, e.g., 1979) the UP is exacerbated but not changed. I argue that the nature of perceptual error is akin to cognitive error when it comes to the more complex and meaning-laden percepts of everyday perception, as opposed to line drawings that are typically referred to in the context of illusions_m . Just as with perceptual errors, cognitive errors often do not become manifest. However, if they do become manifest, they can typically be corrected with much greater ease than can perceptual illusions_m , which may well be the distinguishing feature between perceptual and cognitive error. Cognitive errors become noticeable more indirectly by recurring to a short-term memory of a dissenting fact or by reasoning—which is often faulty by itself. The literature about cognitive error is enormous. To give one classical example, we have trouble with simple syllogistic reasoning, in particular if negations are used. Wason's famous selection task (Wason & Johnson-Laird 1972) shows how limited our abilities are (Figure 9). Imagine you have four envelopes in front of you. You are to test the statement “if there is sender information on the back side then there is a stamp on the front”. Which of the 4 envelopes do you have to turn over? Do not turn over any envelope unnecessarily.

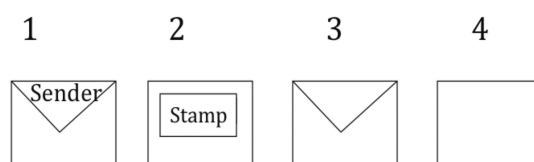


Figure 9: Which envelopes do you have to turn to test the statement “If there is sender information on the back side then there is a stamp on the front”?

Well—it is easy to see that envelope 1 has to be turned (modus ponens), but then it gets harder. Many observers think that envelope 2 needs to be turned. However, this is not the case. Only 4 has to be turned in addition to 1.

A sender on its back would violate the rule (modus tollens). The majority of college students fail to solve this problem, but as soon as the context is changed, all mistakes can be eliminated. In the context of screening for drinking underage, all observers perform accurately (see Figure 10). Here again 1 and 4 need to be “turned over”. Only by thinking the problem through or by noticing that the problem structure is identical to the envelope scenario and the wine drinking scenario does the error become manifest. We may or may not want to call it a cognitive illusion. This term is not widely used for such mistakes or fallacies, with the exception of Gerd Gigerenzer and his research group (see e.g., Hertwig & Ortmann 2005). However, even if we call these mistakes cognitive illusions, they are different in nature from perceptual illusions_m (which typically contain a judgmental aspect). We do not readily notice cognitive illusions. Although the distinction between perception and cognition has outlived itself (and cannot be made with clarity to begin with, see above), for practical convenience, I will continue to use the terms to emphasize cases where deliberate thought processes enter the equation. We happily live with many a fallacy without ever noticing. Millions went through their lives believing in impetus theory and seeing the sun circle around the earth, let alone holding seemingly absurd beliefs about the shape of our planet.

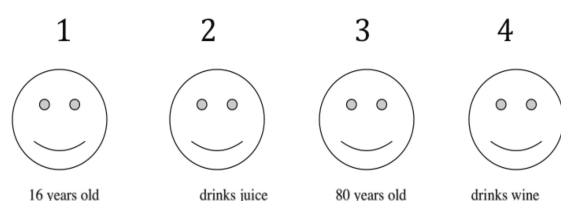


Figure 10: Whom do you have to query about age or beverage type to test if “Only adults have alcoholic beverages in their glass”? It is obvious that the juice drinker and the elderly person need not be queried.

Errors only turn into illusions_m when we become aware of them and at the same time cannot correct the error (easily). Just try to see the earth rotate rather than see the sun rise. It is impossible. We continue to see the sun rise

above a stable horizon, never the other way around. And we continue to misjudge implication rules or widen the grasp of our fingers a tad more when reaching for an Ebbinghaus stimulus even if we know about the illusion (see Franz et al. 2000). Other errors can only be spotted when large data samples are collected and analyzed statistically. For instance, to expert golfers, the putting hole on the green looks larger than it does to novices (Witt et al. 2008; Proffitt & Linkenauger 2013). They will never become aware of this fact, although the fine-grained scaling of perception as function of skill might be functional during skill acquisition. Spectacular as they may be, such errors of which we are unaware should not be called illusions_m because almost all our perceptions and cognitions contain some degree of error. We may believe that a rolling ball comes to a stop because it has used up its impetus, or we may hold that we should aim where we want a moving ball to go rather than using the appropriate vector addition to determine where to aim. As long as our action results do not force us to reconsider, our convictions will remain unchanged. One could say that we have a model of the world, or its workings, that suffices for our purposes.

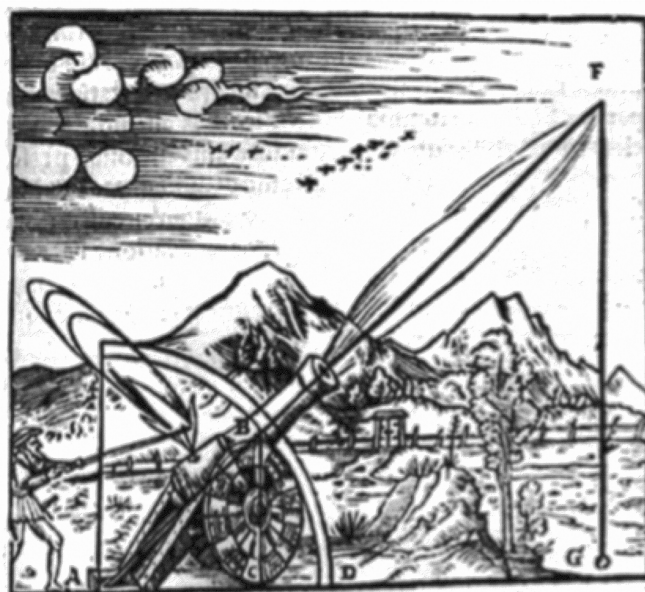


Figure 11: Technical illustration explaining the trajectory of a cannon projectile by Daniel Santbech (1561): *Problematum Astronomicorum*, Basel.

Why are so many researchers willing to call a small manifest discrepancy between two percepts of the same object an illusion, while gross deviations of perception or conception from physical reality are not deemed to deserve the same name? Take the straight-down belief (not illusion). Many observers take an object that is being released from a moving carrier to fall straight down rather than in a parabola (McCloskey et al. 1983). Figure 11 illustrates this belief as it was state-of-the-art physics knowledge from Aristotle through the Middle Ages. It persists today in cognition and perception. Even when impossible events of straight down trajectories are shown in animated movies, to some observers they look better than do the correct parabolas (Kaiser et al. 1992).

Note that there was a discussion at the time whether or not the transition from the upward impetus to the downward impetus was immediate or if a third circular impetus inserted itself, such that there were be two trajectory changes. The intermediary could only be thought of as linear or as a circular arc—anything else would have been too far from divine perfection. Presumably, the more principled physicists before Galileo favored the simple transition. Others, such as Aristotle himself, presumably preferred the interstition of the circular arc, as it would reconcile trajectory observation with the physics of the time. The pre-Newtonian thinking about projectile motion nicely illustrates that we see the world as in accord with our actions. To the medieval cannoneer, what he saw and understood about projectiles was sufficiently accurate, given the variance introduced by the inconsistent quality of the gunpowder and the fluctuation in the weight of cannon balls at the time.

Thus, we have argued that visual illusions_m, just as cognitive illusions_m, have to become manifest to be called such. They are a special and rare case in which the discrepancy between a percept and what an ideal observer should have seen instead is noticed. Normally this discrepancy goes unnoticed. We will now take a look at why it goes unnoticed and argue that an illusion_d will only alter perception if it interferes seriously with our action requirements. As the latter vary among

people, illusions_d can be private and may be very far from the truth—as, for instance, in the context of projectile motion (see Hecht & Bertamini 2000). The private aspect of perception is to be taken as unconscious in the sense of Helmholtz. For instance, we do not only think that a baseball thrown toward a catcher will accelerate after it has left the thrower's hand (which may even be incompatible with impetus theory), but doctored visual scenes in which the ball does accelerate are judged as perfectly natural looking. This amounts to the perceptual analogue of what Herbert Simon (1990) has called satisficing in the domain of reasoning and intuitive judgment. The visual system searches until it has found a solution that is satisfactory, regardless of how far away it is from a veridical representation of the world.

To conclude this section, we believe that perception of objects, be it the stick in the water or a falling brick, is a solution to the underspecification problem. Perception is always fraught with error in the sense of a discrepancy between the percept and the underlying physics. This error only becomes manifest when a simple perceptual judgment or comparison reveals a contradiction. In all other cases the error goes unnoticed. Two such cases will now be described at length to make the point that perceptual illusion_d is the rule rather than the exception.

2 Two case studies or how we deal with error

The study of geometric illusions or overestimation of slope, distance, and size as a function of situatedness misleads us into believing that perception normally reveals the true state of affairs. The finding that golf holes look slightly bigger to experts as compared to inexperienced golfers is spectacular because and only if we assume that perception is normally veridical. This is, however, not the case. Normally, our grasp of the physical world is rather limited. I present novel data from two everyday domains that differ from the standard examples of intuitive physics in a crucial way. They deal with the understanding (first case study) and the perception (second case study) of relational properties, rather than with more straight-forward perception of simple properties.

Seeing the color of an object or its size, predicting its motion trajectory, etc., refer to simple properties. Most everyday activities, however, involve relational properties. We need to see and predict how we might interact with objects in the world. This interaction depends on our own makeup, on the object's properties, and on the relation between the two. For instance, to judge whether a slope might be too slippery for us to walk on depends on the quality of the soles of my shoes, the surface texture of the slope, and also on their interaction. A polished hardwood ramp may be slippery if I am wearing shoes with leather soles, but it may be very sticky if I am barefoot.

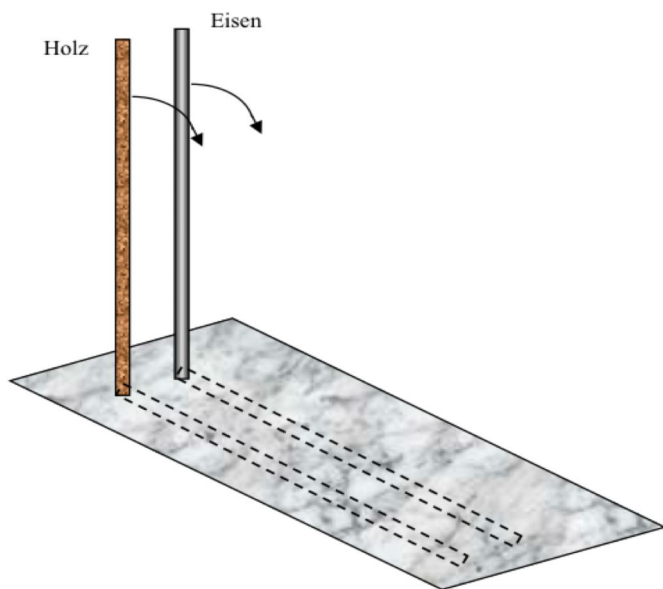


Figure 12: Task 1: The rod on the left is light, it is made of wood; the rod on the right is heavier because it is made of iron. If they begin to tip over at the same moment in time, which one will fall faster?

The two case studies that follow are intended to illustrate in detail how limited our understanding of relational properties is in general, and to show that we have to make decisions in the face of poor perception that may have serious consequences.

2.1 Case study: Balancing as a relational property

Before you read on, please take a minute to solve six questions about the depicted falling rods. Solutions will be provided later. Note that in tasks 1

through 3 (see Figure 12, 13, 14), the scenario is as follows. Two rods are held upright, but they are very slightly tipped to one side (by exactly equal amounts), such that they will fall once released. They are released at exactly the same moment. Which one will hit the ground sooner? In tasks 4–6, you are to judge the ease of balancing such a rod on the tip of your index finger.

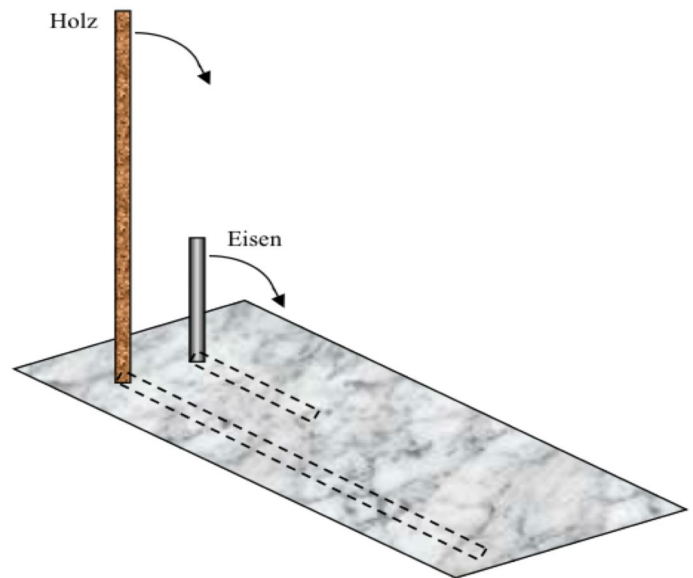


Figure 13: In Task 2 the rods are equally heavy but have different lengths. The left rod is made of wood; the rod on the right is shorter but has the same weight as it is made of steel. Which one will fall faster?

Task 4 asks about the same rods as in Task 1, but the question is whether the wooden or the steel rod would be easier to balance on the tip of your index finger.

Task 5 asks whether the short steel rod or the longer wooden rod of equal weight would be easier to balance on the index finger. And finally, Task 6 asks whether a weight attached to a given rod would make it easier to balance, and if so, where it best be attached (top, center, bottom). In a large survey, we tested the intuitive knowledge of a large number of college students about these tasks. Note that we tested such that each subject only had to solve one of the six tasks.

2.1.1 Methods detail

180 college students (123 women, 57 men, age $M = 24.9$ $SD = 5.9$, ranging from 18 to 53

years) volunteered to participate in the survey. We used a paper-and-pencil test to investigate the subjects' knowledge and to obtain their estimates about which objects would be easier to balance. The six tasks were explained carefully and illustrated with drawings similar to those shown in Figure 12, 13 and 14.

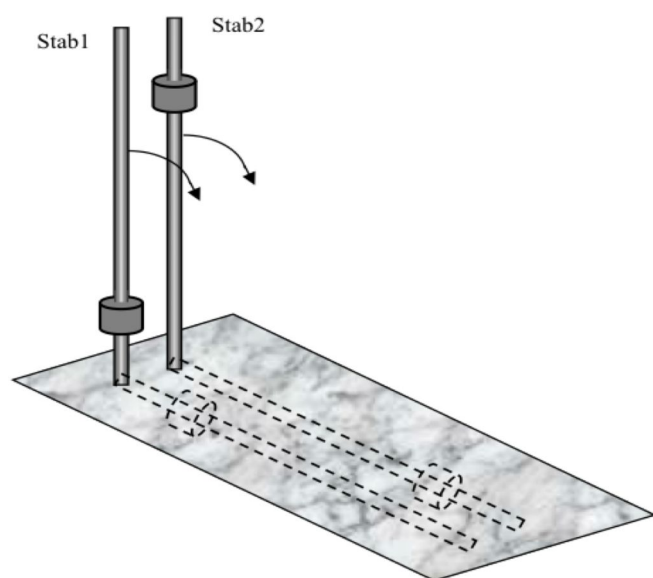


Figure 14: In Task 3, the two rods are identical in material, length, and weight. An additional weight is attached either at the bottom or at the top. Which rod will fall faster?

Each task was presented to 30 students. Tasks 1–3 were used to test intuitive knowledge without referring or alluding to the act of balancing. Merely the process of falling from an almost upright position to a horizontal position had to be judged. In the first task (Figure 12), subjects saw two rods of equal length (1m) but of different material and weight. The wooden rod was said to weigh 40g, the steel rod 400g. The accompanying information text indicated that both rods were slightly tipped over at the exact same time, for instance by a minimal breeze. The wooden rod was to take exactly 1.5 seconds to fall from its upright position to the horizontal. We had tested the falling speed of such rods and measured it to be approximately 1.5s. The subjects were asked to estimate the fall-duration of the steel rod. The second (Figure 13) task showed two rods of equal mass (40g) but different length (rod 1 = 100cm, rod

2 = 36cm). The information text was the same as in Task 1. The third task (Figure 14) showed two rods of equal length (1m) and weight (220g). However, an additional small object (220g) was placed respectively toward the top or the bottom of the rod (rod 1 = 10cm from the bottom, rod 2 = 90cm from the bottom). The accompanying information text indicated that both rods would be tipped over by a minimal breeze and that it took rod 1 exactly 1.5 seconds to fall to a horizontal position. Subjects were to estimate the fall-duration of rod 2.

Tasks 4–6 used the same rods but the questions about them were couched in the context of balancing. This should evoke experiences that subjects may have made when balancing or hefting objects. Thus, rather than asking which rod would fall quicker, we asked which would be easier to balance.

The fourth task showed the same two rods of equal length (1m) but different weights (wooden rod = 40g, heavy steel rod = 400g) that had been used for Task 1 (Figure 12). The subjects were asked to indicate which rod they thought they could better balance on the tip of one finger, typically the index finger. The possible answers ranged from 1 (“rod 1 much better than rod 2”) to 7 (“rod 2 much better than rod 1”). The fifth task (Figure 13) showed two rods of equal weight (40g) but different length (rod 1 = 100cm, rod 2 = 36cm). Again, the subjects were asked to indicate which rod they could better balance with one finger. Task 6 showed one rod (length = 1m, weight = 220g). Subjects had to indicate the position that they would place an additional small object (mass = 220g) to get optimal balancing characteristics (from 10cm = bottom to 100cm = top). It was made clear that the weight would not come into contact with the balancing hand even when it was placed at the bottom.

2.1.2 Results

People who cannot draw on formal physics training to answer the six tasks have a rather poor intuitive understanding of falling rods. Neglecting air resistance, the rate of falling is determined by how high the center of gravity

(barycenter) is above the ground. The rod's mass is irrelevant. Thus, rods of equal length (mass distribution is assumed to be uniform) fall at the same rate, but the shorter rod falls quicker than its longer counterpart. By the same token, a weight attached to the tip of the rod should cause it to fall more slowly because it moves the barycenter closer to the tip.

In general, the subjects estimated their knowledge in the natural sciences to be moderate when asked to judge it on a six-point scale ranging from very poor (1) to very good (6). Mathematics knowledge ($M = 3.62$, $SD = 1.13$) was judged better ($t(179) = 11.98$, $p < .001$) than physics knowledge ($M = 2.56$, $SD = 1.26$). The men estimated their knowledge somewhat higher than did the women, for physics ($t(178) = 8.8$, $p < .001$) and mathematics ($t(178) = 2.34$, $p < .05$).

Task 1: In reality, both rods fall with the same speed, as Galileo Galilei showed in 1590 with the help of several experiments about free fall (e.g., [Hermann 1981](#)). The falling speed is independent of their mass as long as air resistance is negligible. Thus, 1.5 seconds was the right answer. 40% of the test subjects answered correctly. 43.3% estimated that the heavier rod would fall faster, while 16.7% estimated that it would fall more slowly.

Task 2: Because of the lower barycenter the shorter rod falls faster and its fall-duration is briefer than 1.5 seconds. 46.7% of the test subjects indicated this. 44.3% thought that the fall-duration would be the same and 10% estimated that the shorter rod would fall more slowly.

Task 3: Because of the higher barycenter, the second rod falls more slowly. Therefore, its fall-duration is longer than 1.5 seconds. Only 20% of the subjects chose the right answer. 50% estimated that the rod with the higher barycenter would fall faster and 30% estimated that it would fall at the same rate.

There is a direct link between the fall-duration of an object and the ability to balance this object. The longer the fall-duration, the more time there should be to move the balancing finger right underneath the barycenter, and hence the easier to balance (a moderate weight

assumed). We confirmed this hypothesis empirically in several experiments where subjects actually had to balance different rods to which weights were attached at different heights. Thus, we can predict the ability to balance different objects by comparing their fall-duration.

Task 4: Here, the rods (same length, different weight) had the same fall-duration—so the ability to balance them can be assumed to be the same, too. This was recognized by only 3.3% of the test subjects, while 73% favored the heavier rod, and 23.3% the lighter one.

Task 5 (two rods, same weight, different length): Because of the longer fall-duration the longer rod is easier to balance. This was assumed by 56.7% of the subjects. 20% estimated both rods to be equal and 23.3% thought the shorter one would be easier to balance.

Task 6 (additional weight): The higher the barycenter the longer the fall-duration—and with it the ease of balancing. Therefore, the additional object should be placed at the top of the rod. This was indicated by 33.3% of the test subjects. The majority of 43.3% chose the bottom for placing the object, and 23.3% chose positions between bottom and top.

In sum, the intuitive knowledge about the fall of different objects is rather spotty. About half of the subjects knew that fall-duration is independent of mass (Task 1) and that shorter objects fall faster (Task 2), only 20% realized that the position of the barycenter is relevant and that the fall-duration increases when the barycenter is shifted to the upper end of the rod (Task 3). This is remarkable because on a daily basis we handle objects whose barycenter differs from the geometrical center, for instance a filled vs. an empty soup ladle, top-heavy tennis rackets, etc.

Asking directly about the act of balancing did not reveal superior understanding. When asked about their ability to balance objects, people do know that longer objects are easier to balance than shorter ones, but they do not seem to realize that the mass of the object is irrelevant (Tasks 4 and 5). In other words, although a majority of our subjects was able to recognize that mass is irrelevant for fall duration, they failed to see the irrelevance of mass in the rela-

tional balancing task. The involvement of the own motor action appears to have made the judgment task more difficult. The important role of the position of the barycenter (i.e., mass distribution, Task 6) went equally unnoticed in the falling and the balancing tasks. In general, knowledge about balancing properties and the underlying physical principles can be described as rather moderate. Do experts have a superior understanding of these principles?

2.2 Extending the case study: Comparing physics experts with non-experts

As all subjects had judged their physics knowledge to be rather limited, we chose to test a group with formal physics training on the balancing questions. We also tested a social science control group and added two new tasks. Tasks 1–3 were dropped from the study, while Tasks 4–6 were included. To test for a specific heuristic, namely that heavy objects are harder to balance, the following two tasks were added:

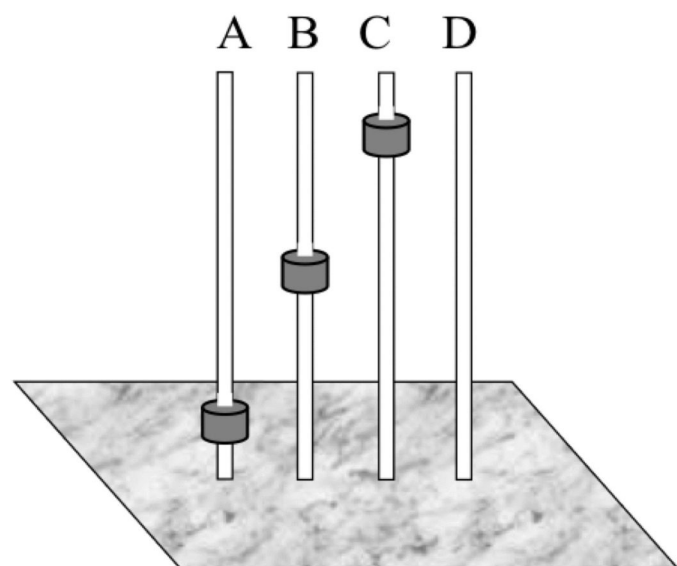


Figure 15: In Task 8, the four rods labeled A–D should be sorted according to the difficulty of balancing them. The correct order is C–B–D–A or C–D–B–A.

Task 7: The question “Does a weight help and if so, where would you place it?” was posed with respect to the much lighter wooden rod ($m = 40\text{g}$). Thus, Task 6 was replicated with a lighter rod. Finally, a more fine-graded question

was added to assess by how much expert knowledge would be superior to normal knowledge, if at all:

Task 8: The eighth task showed four rods of the same material (steel, length 90cm). On three of them, a weight was attached at different positions (as shown in Figure 15). The subjects had to order them according to which would be easiest to balance on the tip of one finger. Note that the height of the barycenter matters. It is equally located in the center of rods B and D.

2.2.1 Methods detail

Participants: 84 college students, mainly of Psychology (69 women, 15 men, age ranging from 19 to 66 years) and 113 college students of Physics, Mathematics, and Chemistry (41 women, 72 men, age ranging from 18 to 27 years) were tested. The students of mathematics, physics, and chemistry estimated their knowledge in mathematics ($M = 2.65$, $SD = 1.02$) and physics ($M = 2.68$, $SD = 1.07$) to be moderate. The men estimated their knowledge of physics to be higher than did the women ($t(111) = -4.34$, $p < .001$). No difference was found for self-assessed maths skills ($t(111) = -.22$, $p = .83$).

A paper-and-pencil test was used to investigate the assumptions subjects held about the effect of various object properties on how easily the respective rods could be balanced. The test booklet included eight tasks: one per page. Each task consisted of a hypothetical scenario illustrated by a drawing. Different pseudo-random orders of the eight tasks were executed by all students. Tasks that built upon one another were kept in their logical order. Once a given task was finished, the page had to be turned. It was not permitted to go back to a previous page. Depending on the task, subjects had to make a binary choice (pick one or answer yes or no) or they had to grade their answers on a seven-point scale, according to how sure they were that one alternative would win over the other (certain win, very likely, somewhat likely, equal chance, somewhat unlikely, very unlikely, certain loss).

2.2.2 Results and discussion

Task 4 (equal length, different weight of the rods): Only 3.6 % of all social science students produced the correct solution and stated that the wooden and the steel rod would be equally hard to balance. Half of them thought that the steel rod would be easier to balance, and the remaining subjects chose the wooden rod. This corresponds well to the results obtained with the first large student sample. The physics students, in contrast, performed better albeit nowhere near perfection. 22% of them chose the correct answer. 21% thought the wooden rod would be easier to balance, and 57% thought the steel rod would be easier to balance. Thus, social scientists equally chose one or the other whereas physicists preferred the metal rod, and at most one fifth of them knew the correct answer (provided they were not just guessing better than the social science students).

Task 5 (equal weight, different length): Half of the social science students (53%) correctly thought that the longer rod would be easier to balance, and less than 2% thought that length did not matter. The physics students did noticeably better: 76% chose the longer rod, and only 20% thought the shorter rod would be easier to balance. 4% thought it would be the same with both rods.

In **Task 6** (attach weight to steel rod): 60.7% of the social scientists thought that a weight would make it easier. When asked to place the weight, only 9.5% put it in the top third (for analysis purposes the rod was divided in three equal parts), and 44% placed it at the bottom third. Physics students fared a little better. A mere 44 % thought that a weight would improve balancing, but those who did correctly placed the weight at the top (40% of all physics students).

Task 7 (attach weight to light wooden rod): Not surprisingly, performance was very similar to Task 6 ($r = 0.76$). If anything, the rod's being lighter improved performance. 77.4% of the social scientists thought that a weight would make it easier. When asked to place the weight, only 19% put it in the top third. 45.2% put the weight in the bottom sec-

tion, and the remaining students placed it in the middle section of the rod. Physics students fared a little better. 81% thought that a weight would improve balancing. However, the correct placement at the top was made by only 40%. Thus, in light of the results from Task 6, it seems that those who knew the correct answer were unimpressed by the weight of the rod. However, among those experts who merely guessed and suspected that weight would make a difference, they guessed so more often when the rod was lighter—increasing the salience of the weight.

Task 8 (order the rods): Social science students: According to the reasoning that a greater moment of inertia should facilitate balancing (note that this will not hold for much heavier rods), the correct order is C, B = D, A. Not a single subject produced this answer. 16.7% chose the order A, B, D, C; another 16.7% chose A, D, B, C. Only one subject considered a tie, albeit with a wrong ordering (B, D, A=C).

Physics students: Notably, 6% of the subjects did give the correct answer of CBDA or CDBA. 94% of the subjects answered incorrectly. Thus, the physics students were somewhat more knowledgeable than the social science students.

In sum, the errors we make in perceiving the balancing properties of simple objects are large. The important variable of mass distribution is ignored entirely. We plainly do not see how an object is best balanced until we try it out, even though we balance objects on a daily basis. Most if not all observers are unable to correctly imagine or remember past balancing acts. Formal physics training has surprisingly little effect on the paper-and-pencil task for assessing falling and balancing of rods. Note that the classical mechanics knowledge that would help solve the problem should have been held by all natural science students involved in the study. The fact that their answers were only slightly superior to novice intuitions is stunning. Why is the textbook knowledge of classical mechanics so frail that it has not been internalized, such as to inform our intuitive judgments or at least facilitate our textbook learning?

Throughout evolution we had to handle and wield objects by balancing them. One might argue that such knowledge is not available to the ventral processing stream (see [Milner & Goodale 2008](#)). However, in further tests we confirmed that performance in our tasks did not improve when we let subjects wield a rod before filling out the questionnaire. Even though observers are able to feel how long a stick is when wielding it while being blindfolded (see [Turvey & Carello 1995](#)), they are unable to exploit the available perceptual cues that inform them about the balancing properties of an object. Thus, although we know shockingly little about balancing, it seems to be sufficient to guide our daily actions. We correctly see longer sticks as being easier to balance, but we fail to see the importance of mass distribution. Even when educated by formal physics training, our performance becomes only slightly more sophisticated. The gap between percept and reality closes merely by a small amount. It appears that the visual systems of different observers adopt different private models that often include rod length but not mass distribution.

2.3 The second case study: Visual cues to friction

Let us now look at another relational property that may have more serious consequences for our health: friction. If we misbalance an object, we may break it, but if we misjudge the slipperiness of the surface we walk on, we may get hurt. We need to avoid accidents on slippery ground and we have to estimate the force we need to apply to hold an object. Importantly, we often cannot wait for haptic cues to make this information available, but typically we have to make the underlying judgment of slipperiness on the basis of visual cues. The mere look of a wet slope may be all we have to guide our decision to tread forcefully or to hold on to a hand-rail and walk gingerly. The human ability to make such visual assessments of slipperiness is not well explored. We hold that this is because friction is not a simple surface property but rather a relational property, which can only be determined by relative characteristics of two

surfaces. In other words, the fact that a surface is rough does not imply high friction, and the fact that a surface is smooth does not imply that it is slippery. Plastic for instance, can be very sticky on human skin but very slippery on wool or felt.

In what follows, we provide an overview of friction perception and briefly introduce venues to visual and haptic roughness perception. Then we report two experiments that were conducted to assess visual and haptic judgments of friction between surfaces.

2.3.1 Friction as a relational property vs. surface roughness

Some surfaces afford walking on whereas others do not. The information that allows the organism to make potentially critical decisions about where to tread or how strong a grip should be is based on a variety of perceptual dimensions (see e.g., [Michaels & Carello 1981](#)). Even when ample opportunity is given to haptically explore the surface, its felt roughness is not necessarily the same as the friction between the exploring hand and the surface, let alone the friction between the sole of the shoe and the surface. For instance, if our hand is moist we feel high friction when exploring a polished marble floor and at the same time we feel it to be very smooth. We may even perceive it as slippery—factoring in the effect of dry vs. moist hands.

Tactile competence regarding perceptual access to roughness of surfaces appears to be rather sophisticated (for a state-of-the-art review of haptic perception see [Lederman & Klatzky 2009](#)). In essence, haptic perception of surface roughness is better when the surface is explored dynamically as opposed to statically. Errors are generally rather small. More interestingly, several studies have demonstrated that cross-modal sensory information (e.g., vision and touch) can lead to better estimates of a texture's roughness (e.g., [Heller 1982](#)). Other research has also shown that different sensory modalities are weighted about equally when estimating the roughness of textures ([Lederman & Abbott 1981](#); [Lederman et al. 1986](#)).

Even by mere visual inspection, observers are able to see how rough a surface is (Lederman & Klatzky 2009). Such findings may have tempted researchers to unduly reduce friction to surface roughness. For instance, in the ergonomics context of accident analysis, slipperiness of work surfaces is typically operationalized by surface roughness, with the implicit or explicit assumption that roughness is good enough an approximation of friction (see e.g., Chang 1999; Chang et al. 2001; Grönqvist et al. 2001). However, friction is a rather complicated property between surfaces, for one because it is subject to change with the amount of pressure one applies or with the speed at which the surfaces move relative to one another. And people appear to have some difficulty judging friction (Joh et al. 2007).

Let us consider the case of a square block of cement on a large wooden surface. The heavier the block, the higher the friction coefficient. And the rougher the surface of the block the higher the friction coefficient. Thus, friction is a function of the force applied to a given surface, of area, and of roughness. Children and adults seem to be able to perceptually appreciate some but not all of the above-mentioned three components of friction. This intuitive knowledge develops with age. Adults have some insight into the multiplicative relation between the weight of an object and its surface texture in cases where the object is pulled across a surface, whereas nine-year-old children seem to assume a simpler additive relationship (Frick et al. 2006).

Friction is defined by the interaction between two surfaces, and its estimation requires knowledge about how different surfaces can interact. Thus, the seemingly simple visual percept that we have of a surface as “slippery” is a rather complex physical relation that pertains between properties of the surface and the contact object. Physically, slipperiness is indicated by the friction coefficient between two surfaces, which is usually measured by placing an object on an adjustable ramp. As the steepness of the ramp increases, one determines the angle at which the object starts to slip (static friction) or when the object starts to move uniformly (kinetic friction).

We can haptically judge the roughness of surfaces, and we are also able—to some degree—to haptically judge the friction between surfaces. Grierson & Carnahan (2006) have shown that individuals can haptically perceive slipperiness; that is estimates were significantly correlated with the friction coefficients between an object’s surface and skin. In their first experiment, they showed that tangential motion is required to judge the friction coefficient realistically. In a second experiment, they examined the force people applied to lift an object with a certain weight and surface structure. The applied force was often higher than necessary. Next to nothing is known about our ability to judge slipperiness based on visual information.

2.3.2 Slipperiness Experiment: Visual cues to friction of familiar surfaces

Vision has been shown to improve haptic judgments in endoscopic surgery. Within a simulated endoscopic environment, Perreault & Cao (2006) tested the effects of vision and friction on haptic perception by measuring for how long participants held on to the objects with the surgery tool. In a second experiment, participants had to compare the softness of pairs of simulated tissue. The experiments showed that visual and haptic feedback were equally important for the task. This suggests that visual cues can be exploited to judge slipperiness.

Presumably the main visual cue for predicting slipperiness or friction is shine (gloss, reflection, etc.) of a surface. Joh, Adolph, Campbell & Eppler (2006) explored which visual information can serve as a warning of low friction surfaces. They asked their participants which cues they use to identify slippery ground, and tested whether visual information is reliable for the judgment of slipperiness under different conditions (indoor and outdoor lightening). Walkers seem to rely on shine for selecting a safer, less slippery path, even though shine is not a very reliable visual cue for indicating slippery ground.

With two experiments we attempted to assess, in more general terms, the ability to perceive slipperiness. In our first experiment, we tried to

find out to what extent visual and haptic information enables us to estimate friction between two surfaces and, in particular, how far visual cues in isolation decrease the ability to judge friction. In our second experiment, we manipulated the visual appearance of given surfaces to explore the effects of glossiness, contrast, and undulation on perceived friction.

Every day we encounter different types of surfaces with which we are in contact. In these situations we do not really think about how much force is to be exerted in order to create sufficient friction, be it between the fingers and the object we are grasping or between the sole of our shoes and the surface of the road we tread. Nonetheless, we rarely accidentally drop an object or slip on the road. Thus, we must have some degree of intuitive knowledge about the friction of surfaces. The experiment sought to find out, first, if this is really the case, and then which sensory information might guide our estimates of friction.

2.3.3 Methods detail

33 female and 31 male subjects between 18 and 52 years of age ($M = 25.3$; $SD = 6.6$) volunteered and were paid for participating in the study. All had normal or corrected to normal vision, and no one reported haptic impairments.

Ten different types of surfaces (see Figure 16) were glued onto separate thin quadratic tiles of wood with a size of 10 x 10cm. The surfaces were sheets of Teflon, pan liner, smooth and rough foam rubber, cloth, felt (soft and hard), and three different grades of sandpaper.

Two common reference surfaces were picked: human skin and smooth untreated wood. That is, the participants had to estimate the friction of the above ten surfaces with respect to one or the other of the two reference surfaces, skin or wood.

To measure the perceived friction, a ramp was used. Its slope could be adjusted to a steepness corresponding to the setting where the tile was judged to start sliding down. The ramp consisted of two wooden boards connected with a hinge. It was placed in front of the participant and could be continuously adjusted (see Figure

17). A measuring stick was attached to the top of the ramp such that the experimenter could easily record the height of the ramp while the participant saw only the unmarked side of the measuring stick. The height settings were then converted to slope angle, which in turn was used to determine the friction force acting between ramp and probe surface.



Figure 16: The ten materials used in the first experiment. Top row from left to right: Teflon, pan liner, smooth and rough foam rubber, cloth. Bottom row from left: felt (soft and hard), three different grades of sandpaper (320, 180, 40 in that order). All materials were mounted on identical square wooden tiles. The matchstick is shown to provide scale information, it was not there in the experiment.

The slope of the ramp used to estimate the friction of the different surfaces could be varied from 0 to 90 degrees. We computed coefficients of estimated static friction for the subsequent analyses using the following equation:

$$\mu_H = \mathbf{F}_R / \mathbf{F}_N \text{ (friction coefficient = friction force / weight)}$$

A 4 x 2 x 10 design was used, with one four-level between-subjects factor (Condition), and two within-subject factors, Reference Surface (two levels: skin and wood), and Surface Material (ten levels: Teflon, pan liner, smooth and rough foam rubber, cloth, felt (soft and hard), and three different grades of sandpaper).

The factor Condition consisted of different instructions for exploring the surface materials (see Table 1). In the haptic-visual condition, observers were asked to touch the surfaces and to visually inspect them. In the haptic condition, the surfaces were hidden in a box at all times and could only be explored haptically. In the visual condition, observers were not allowed to

touch the surfaces but could inspect them visually. In the photo condition, finally, observers merely viewed photographs of all ten surfaces. The same photographs as depicted in Figure 16 were used, with the exception that the match was not present. The photographs were the same size as the actual tiles (10 x 10 cm).



Figure 17: The ramp used to measure the estimated friction coefficients produced by the participants. The ramp had to be adjusted to the angle at which the respective tile would just about start to slide. In the case of skin as reference surface, observers were told to imagine the ramp to be their torso or to be covered with skin.

Subjects were allowed to look at the respective reference surface (skin or wood) before making a set of judgments based on this reference surface. They were also allowed and encouraged to touch the reference surface regardless of the condition in which they were tested. That is, even the group that could only visually inspect the test surfaces had visual and haptic experience of the generic reference surface. The ramp itself was not to be touched in this phase of the experiment, in order to ensure that the groups did not differ in how they explored the ramp. To envision the friction of skin, subjects were instructed to touch the inner side of their forearm, and to envision the friction of wood, they had a piece of wood (the same wood also used for the ramp) lying in front of them that they could touch. Half of the participants started with wood as reference and then after a short pause used skin as reference. The other half started with skin and then judged wood.

Within each block, the order of the surface tiles was randomized separately for each observer.

Table 1: The four test conditions under which separate groups of subjects were asked to explore the material surfaces.

I. Haptic-visual	The ten test surfaces were visible and could be touched without any restrictions.
II. Haptic	The surfaces were hidden in a box and could only be touched but not seen.
III. Visual	The subjects were only allowed to look at the surfaces.
IV. Photo	Photos of the surfaces were presented on a TFT-display, so only restricted visual information was available.

The procedure consisted of three parts. First, subjects had to estimate the friction of the ten materials, all presented successively and in random order. To do so, they had to adjust the slope of the ramp (see Figure 17). After inspecting the reference surface and the first tile, they had to set the ramp's slope to the point where they expected the particular surface to just start slipping on the ramp. The surface tiles were never physically placed on the ramp. Then the remaining nine tiles had to be judged in the same manner.

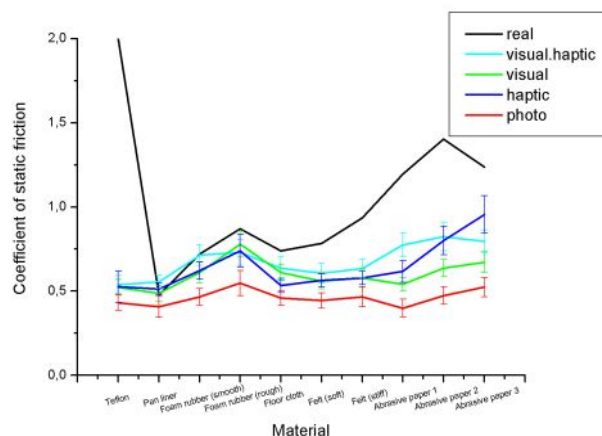


Figure 18: Actual and perceived coefficients of friction between **skin** and the respective materials. The solid black line corresponds to the actual angle of the slope at which the tile would indeed start to slide. The other lines represent subjective judgments averaged across all participants of each group respectively. Error bars indicate standard errors of the mean.

In the second part, a short questionnaire was given to the subjects. Finally, the procedure was

repeated with the other reference surface. The order of which reference surface was chosen first was counterbalanced such that half the observers started with wood and the other half started with skin.

2.3.4 Results

Line graphs show the actual and the averaged estimated coefficients of static friction on skin (Figure 18) and on wood (Figure 19). With the exception of Teflon on skin, friction was perceived, albeit underestimated. In some cases, roughness appears to have guided perception. For instance, the different grades of sand paper produce similar friction because roughness and contact area trade off against one another. The coarse paper is rougher but at the same time provides fewer contact points than the fine paper. The resultant friction is in fact comparable. However, the coarse paper was mistakenly thought to produce more friction than the fine paper.

With skin as reference surface, haptic exploration improved performance but estimates remained far from perfect. Teflon in particular was grossly mis-estimated. The overall results showed significant main effects of Material ($F(5.7, 342.4)=22.85, p<.001$, partial $\eta^2 = .27$) and Reference Surface ($F(1.0, 60.0)=17.80, p<.001$, partial $\eta^2 = .23$). In addition, the effects of Condition were more pronounced for the reference surface of skin; the interactions of Material x Condition ($F(17.1, 342.4)=2.92, p<.001$, partial $\eta^2 = .13$) and between Material x Surface ($F(7.9, 475.2)=2.87, p=.004$, partial $\eta^2 = .046$) were significant. The interaction of Material x Surface x Condition was also significant ($F(23.8, 475.2)=1.69, p=.023$, partial $\eta^2 = .078$). Contrasts revealed that performance was poorer in the photo condition compared to the haptic condition ($p=.023$) and the haptic-visual condition ($p=.007$). The latter two did not differ significantly from one another or from vision alone.

The post-experimental questionnaire revealed that most participants attempted to use all available information and that they tried to find out which material they were confronted with. After identifying the material, they estim-

ated the friction on the basis of their experience. Perhaps some erroneous estimates could be ascribed to such cognitive influences upon friction estimation.

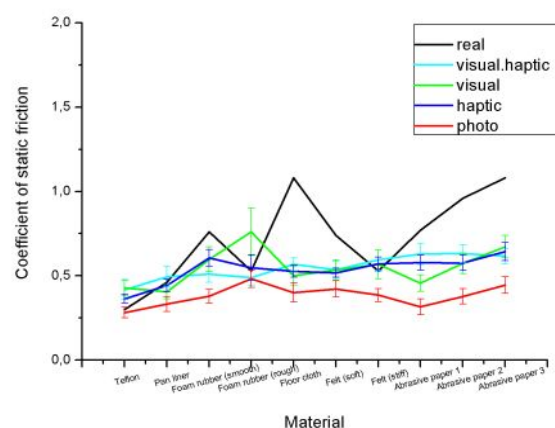


Figure 19: Actual and perceived coefficients of friction between **wood** and the respective materials. The solid black line corresponds to the actual angle of the slope at which the tile would start to slide. The other lines represent subjective judgments averaged across all participants of each group respectively. Error bars indicate standard errors of the mean.

In sum, static friction between a number of different materials and the reference surfaces skin and wood were picked up, but only to a limited degree. Vision alone does transport information about the relational property of friction. This ability to see friction is attenuated but still present when photographs are used. Thus, high-resolution detail appears to be crucial. Surprisingly, haptic cues were not superior to visual cues and even in combination only tended to improve performance. Friction is generally underestimated, with the exception of Teflon and wood, which was grossly underestimated. Multisensory information did not help compared to unisensory information. It appears that multiple information sources improve the perception of simple properties such as roughness (Lederman & Abbott 1981; Lederman et al. 1986), but fail to contribute in more complex cases of assessing friction. When visual information was reduced, not surprisingly, this affected friction judgments negatively. The photo condition produced notable judgment errors. It would

be interesting to find out if this degradation could be compensated for by providing haptic cues together with the photographs. Note, however, that the photographs were able to produce estimates that correlated with actual friction. Thus, some information about roughness is preserved in the photo and can be accessed. The relational property of friction appears to be qualitatively different from and not reducible to roughness.

2.3.5 Friction experiment with manipulated visual appearance

The preceding experiment has shown that observers are able to gain some information about friction by visually inspecting the two involved surfaces together exhibiting this complex property. Given this ability, we should be able to isolate some of the relevant visual surface features upon which this ability is based. In a second friction experiment, we limited the reference surface to wood, and manipulated the visual properties of a select number of surfaces, namely Teflon, foam rubber, and sand paper. Among the changes in visual properties were factors that should influence perceived roughness and thereby potentially also friction, such as convolving the picture with a wave pattern, or changing the contrast in the picture.

2.3.6 Method detail

55 volunteer subjects (23 men and 32 women) participated in the study. They were recruited at the campus of the Johannes-Gutenberg University of Mainz and at a nearby supermarket. All participants were naive with respect to the purposes of the experiment. Their average age was 31.8 years ($SD = 12.6$ and a range from age 16 to 59).

We took some of the pictures of the tiles used previously. The pictures were taken on a Fuji Finepix S5500 digital camera (four megapixels) with a resolution of 1420 x 950 pixels. One reference picture each of coarse sandpaper, structured foam rubber, and Teflon were chosen. Then these reference pictures were modified using four special effects provided by

Adobe Photoshop Six. Five visual effect conditions (Filter) were thus created for each of the three materials (see Figure 20 for the case of sand paper):



Figure 20: The reference picture (n) of the sand paper tile, and the four filter effects applied to the reference picture: ocean effect (o), wave effect (w), reduced lightness (d), and enhanced contrast (c). Note that all pictures were of equal size in the experiment.

1. Normal: The reference picture was the original photo of the surface without any special effect.
2. Ocean: The original photo was convolved with the structure of an ocean surface. A photograph showing the ocean from above with its waves was put as a new layer upon the original photograph with an opacity value of 25%. It added a look reminiscent of structured wood to the photograph. We hy-

pothesized that the added structure would increase perceived friction.

3. Wave: This filter introduced a wave pattern into the picture. This distortion effect was used with the parameters Number of Generators (5), Wavelength (Minimum 10 Maximum 120), Amplitude (Minimum 5, Maximum 35), Scale (horizontal 100%, vertical 100%), Repeat Edge Pixels (On), and Type (Sine). This filter distorts the original structure in the pattern of sine waves. We hypothesized that here, the added structure would not change perceived friction because waves are regular and smooth compared to the ocean texture.
4. Dark: The lightness of the surface was reduced uniformly by 50% (parameter setting: 50). We hypothesized that this would reduce detail, which would decrease perceived friction.
5. Contrast: The contrast was uniformly enhanced such that the according parameter was raised to +50. We hypothesized that the added contrast would emphasize roughness and thereby increase perceived friction.

The photos were printed on high-quality photo paper and shown to the volunteers in succession. The “normal” reference version of one material was always shown first, and then four different versions of the same material were presented in changing pseudo-random orders, for each material respectively. All possible sequences of the materials were presented to different observers. They were asked to imagine the surface shown on the photograph as being the surface of the ramp itself. The same ramp as before was used (see Figure 17), but subjects were not allowed to touch its actual wooden surface. Then they were asked to decide how steep the ramp would have to be set for a wooden tile to start sliding down on the shown surface. The tile of wood was shown to them beforehand and they were asked to touch it. Then they had to put the ramp at the angle at which they thought the wooden tile would just start to slide. As before, we measured the height of the ramp setting in centimetres. With this information, we calculated the angle with

$\sin(\alpha) = \text{height} / \text{ramp length} = \text{height} / 44\text{cm}$ and finally the resulting estimated friction coefficient for all surfaces.

2.3.7 Results

Figure 21 shows the estimated friction coefficients for all three materials averaged across all filters and across the respective reference surface. Friction between wood and foam rubber was judged to be highest, friction with sandpaper was judged intermediate, and friction with Teflon was judged to be smallest. Figure 22 depicts the overall averages by Filter (visual effect). Figure 23 shows the interaction between Material and Filter.

A repeated measurement analysis of variance with Material and Filter as within-subject factors and gender as between-subjects factor was conducted on the judged friction coefficients; F-values were corrected by Huynh-Feldt as necessary. Material had a significant effect on estimated friction ($F(2, 106)=9.54$, $p<.001$, partial $\eta^2=.15$). Foam rubber and paper did not differ, but both were judged to produce more friction than Teflon ($p<.001$ and $p<.003$ respectively).

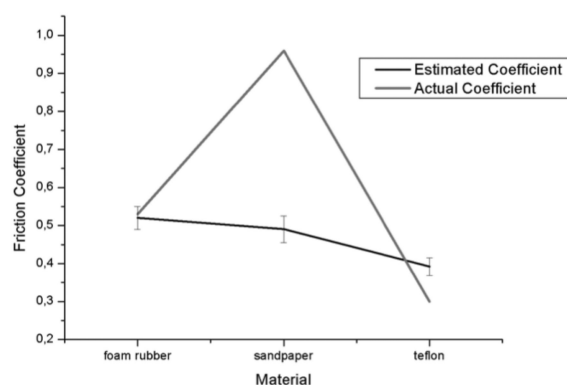


Figure 21: Estimated friction coefficients for the three materials independently averaged across all filters, and the actual coefficients for the three materials on wood. Error bars indicate standard errors of the mean.

The factor Filter also had a significant influence on the estimation of friction ($F(4, 212)=5.351$, $p=.001$, partial $\eta^2=.092$). The unfiltered stimuli were judged to produce the

smallest amount of friction, and all filters appeared to increase the subjective coefficient of friction. Figure 22 shows the estimated friction for all five filters averaged across all three materials. The contrasts between the estimated friction coefficient values for “normal” and “ocean” ($p < .024$), “normal” and “dark” ($p < .001$) and “normal” and “contrast” ($p < .023$) were significant. Because of the sometimes variable judgments, the individual contrasts between “normal” and “wave” as well as “contrast” and “dark” failed to reach significance.

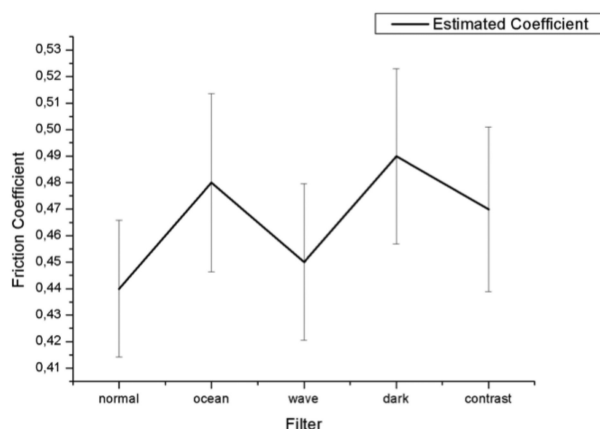


Figure 22: Estimated friction coefficients for the five filters averaged across the three materials. Error bars indicate standard errors of the mean.

We also found a significant interaction between the factors Material and Filter ($F(8, 424) = 3.99$, $p = .002$, partial $\eta^2 = .070$). As visible in Figure 23, this interaction was mainly due to the immunity of Teflon to all filter manipulations and to the special effect of the increased contrast on foam rubber. Now let us have a closer look at the three materials and how they fared with the different filters. Participants could judge the friction between wood and the shown surfaces rather well, with the exception that the friction of sandpaper was underestimated. For some reason some of the grittiness and roughness of sandpaper has been lost in the photos, whereas no such loss occurred for foam rubber and Teflon. To the experimenter, the surface of sandpaper also did not look as rough as it did in real life.

Teflon on wood was clearly judged to be the most slippery surface. Interestingly, the estimated differences between the Teflon reference and its filter-treated variants were very small compared to the other materials (see Figure 23). Presumably, Teflon generally looks so slippery that a ceiling had been approached and the filters could not significantly change the low friction ratings of Teflon. The surfaces that were treated with “ocean” looked like rough wood; the manipulations “contrast” and “dark” seemed to make the structure clearer. The filter “wave” had a smaller influence on the estimations. Participants often said that they found it difficult to classify the wave-treated surface.

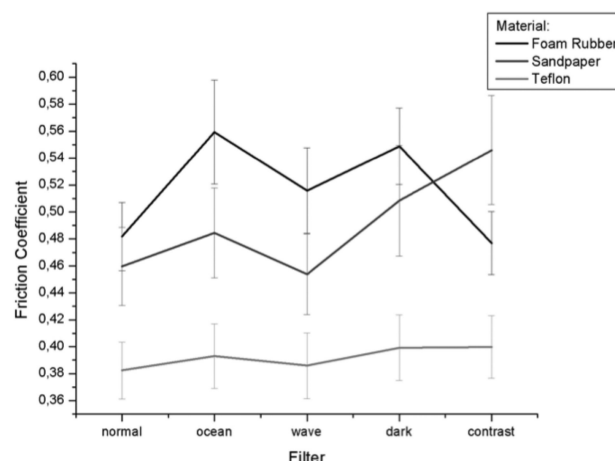


Figure 23: Interaction between the two factors Material and Filter. Error bars indicate standard errors of the mean.

The results of this experiment clearly show that irregular additional structure—as introduced into the surface by convolving the picture with the ocean pattern—causes the perception that the surface is less slippery. This was the case for all surfaces that were not extremely slippery to begin with. Other than hypothesized, reducing the lightness of the surface also tended to produce higher ratings of friction. Increased contrast, on the other hand, produced mixed results. Sandpaper with increased contrast was judged to cause more friction. Contrast had a smaller but similar effect on Teflon. However, when applied to foam rubber, increased contrast had no effect. Taken together, these effects demonstrate that visual aspects of

a surface, such as its microstructure, its lightness, and its contrast co-determine how slippery it is judged to be with respect to a given reference surface. Note, however, that the reference surface was always wood, and simple roughness judgments may have guided the friction estimates.

To summarize the friction case study, we conducted two experiments to assess whether observers are able to visually perceive the complex relational property of friction between two surfaces even when not allowed to touch the surfaces. They were able to do so with limitations. Observers generally tended to underestimate the degree of friction. An underestimation of friction as observed in these two studies could be regarded as a conservative approach to judging the grip force required to successfully grasp objects. Using more force than necessary rarely leads to disaster (consider raw eggs), whereas too little grip force causes an object to slip out of our hand and fall.

The first friction experiment compared judgments based upon visual inspection alone, and then after visual and haptic inspection. Vision in and of itself provides valuable information; additional haptic information added surprisingly little. The second experiment explored the particular visual properties that make surfaces look more or less slippery, but note that the reference surface always remained unchanged. Subjects likely differentiated between surfaces of different roughness insofar as roughness (simple property) and friction (relational property) were correlated. Errors were large in particular when the relational property to be judged was variable. Perceiving Teflon as very slippery (with respect to skin) when it is indeed quite the opposite is a grave perceptual error, but it is not very meaningful to call the error an illusion_d. A perceptual miscategorization of the relational property of friction between surfaces might be a more appropriate description.

3 Conclusion

I have attempted to argue that we need to reconceive the notion of what an illusion is. In the context of the traditional line drawings used

over a hundred years ago to illustrate the shortcomings of vision, illusions_m have begun to misguide our thinking about normal perception. Illusions_m do not indicate the error-prone nature of visual perception. On the contrary, they tend to be small compared to the many illusions_d that go unnoticed on a regular basis. To illustrate that this is the case, I have used two examples from the domain of complex relational properties. This choice was based on the conviction that perception of everyday objects always necessarily includes judgment (be it in terms of Helmholtzian unconscious inference, or be it in terms of private models that may or may not become transparent to the perceiver). The notion that illusions_m should be of interest because they reveal the workings of how the visual system derives percepts from simple sensations is not useful. It is not useful because an illusion_m only becomes manifest by a comparison process that is at least as fraught with cognition as is the perception of everyday relational properties. We have used the classical stick in the water and the equally classical Ebbinghaus illusion to illustrate that illusions_m only become manifest if a cognitive operation is performed (i.e., a perception-inference-cycle when moving the stick or comparing the circle to a reference circle known to be of identical size).

It is also impossible to investigate illusions as merely phenomenal problems. And it is ill-conceived to limit the study of visual perception to seemingly simple phenomena that end up requiring cognition after all. Perceiving is to make perceptual judgments, be they explicit (e.g., by saying which of two objects is larger), or be they altogether implicit, or merely amenable to consciousness by an act of attention (e.g., by determining hand-aperture when grasping an object). It is thus impossible to investigate illusions as purely perceptual errors. Instead, illusions always have a cognitive component in the sense that they require an act of comparison or inference. This holds for all illusions_m, even if they may not be amenable to consciousness. To take illusions as a discrepancy between what we see and what there is, is doubly mistaken. First, there is always a discrepancy (illusion_d) between a visual percept and the object in the world to

which it refers, namely the stimulus. And second, only in rare and simple cases do we notice this discrepancy (illusion_m). The discrepancy is owed to the underspecification problem (UP), the qualitative information gap between the two-dimensional retinal image and the richer three-dimensional percept. The UP puts the perceptual system in a position from which it has to draw additional information from memory, from inference, or from internalized structures that have been acquired throughout evolution. Such structures have been suggested to include that objects are three-dimensional, that light comes from above, that gravity acts along the main body axis when standing or walking, or that the brightest patch in the visual field is usually “white”. Internalized structures gain particular weight if the stimulus is poor. This is the case when looking at simple line drawings and it is all the more the case when looking at relational properties. The quality of solutions to the UP differ greatly as the function of the task demands, but not necessarily as a function of the complexity of the stimulus. On the one hand, the perceptual system achieves performance that seemingly approaches perfection where precise motor action is required in personal space. On the other hand, in more remote action or vista space (for a very useful taxonomy of space see e.g., [Grüsser 1983](#)) some blatant errors are made. Our perception often defies the most basic laws of physics. More often than not do these errors go unnoticed. To illustrate how crudely our perceptions approximate reality even in personal space, we have explored errors in balancing objects and judging the slipperiness of surfaces. When it comes to these relational properties, our perception falls far from the truth. It appears that the errors tend to be as large as they can be without interfering with the perception-action cycle required for adequate or acceptable action. The evolutionary fine-tuning would minimize error until it is no longer relevant for survival. In this sense, normal perception (i.e., the illusion_d) is a satisficing solution. The magnitude of the perceptual errors many observers make is in the league of errors associated with probability judgments (see e.g., [Kahneman et](#)

[al. 1982](#)) and syllogistic reasoning, as opposed to the much smaller errors typically associated with perceptual illusions_m.

Our perception, just like our cognition, has developed to find solutions to problems that suffice. When reaching for an object, perception is accurate enough not to knock it over but to grasp it (most of the time). When judging a surface, it is accurate enough that we do not slip (most of the time). These examples are noteworthy because they do not relegate perceptual error to remote vista space, where precision would not matter. Toppling over an object or falling on a slippery slope concern us in personal space.

In essence, the UP is solved with remarkable accuracy for simple properties of objects within our domain of interaction. However, as soon as the perceptual properties become more complex and involve the relation between two or more objects, the perceptual system can no longer solve the UP with any degree sophistication that goes beyond the level of medieval physics. But rather than giving up and seeing astounding illusions everywhere, the system degrades gracefully and builds theories that suffice for the purpose at hand. Their deviation from reality is not experienced. These perceptual theories may be thought of as more or less universal tools for upholding a meaningful world (in the sense of [Shepard 1994](#)); however, it might make more sense to think of them as universal tools with a private touch that accommodates individual perception-action requirements. A hockey player or a juggler will for instance have developed private models, be they unconscious or amenable to introspection, about friction or balancing that are more sophisticated than the layperson’s. Note that these models need not be explicit, in the sense of a perceptual process, of which the cognitive elements cannot be separated out.

Such private adjustments and elaborations when solving the UP need not be made in the case of classical geometric-optical illusions_m. I hope the above examples and case studies have shown that illusions_d, such as the Luther illusion, do not require detection, and illusions_m that become manifest, such as the Ebbinghaus illusion,

can be upheld because their limited magnitude makes them irrelevant for action.

This raises the questions why illusions_m arise at all. Illusions_m might arise as mere epi-phenomena or as meaningful warning signs for the system to signal that a perceptual fine-tuning is needed. The epi-phenomenon interpretation would suggest that the juxtaposition of two contradictory percepts is a fluke and happens per-chance every once in a while. Optical illusions_m are merely collections of such flukes. The warning-sign interpretation would see in them the purpose of fine-tuning the perceptual system. If the perceptual system subserves action, it would ideally minimize error (illusions_d), and one mechanism to do so would be the experience of illusions_m. It is unclear, however, why illusions would have to become conscious for this fine-tuning to work. Would the necessary re-direction of attention require the experience of an illusion_m? Be this as it may, the system does not even notice error—let alone attempt such fine-tuning—when it comes to perceiving relational properties. Even an approximate veridical perception of relational properties is out of reach of the perceptual system. The system merely arrives at the first solution that satisfies our action needs. A flashy epi-phenomenon or a warning system, as indicated by manifest illusions_m, is not useful here, as the discrepancy between percept and reality is too large.

Now, one might ask about cases where the error is exceedingly large and a warning may indeed be in place. These cases are rare; but they do, however, result in manifest illusion_m, and hence are compatible with the purpose of illusion_m that we suggest. Take for instance the perception of pain in a phantom limb. Here the sufferer does notice the illusion_m. How can pain be so vividly felt in a limb that is no longer there? The warning function of this manifest illusion_m is obvious. For instance, learned reflexes involving the absent limb need to be extinguished and reprogrammed. A more interesting case is the infamous rubber-hand illusion (Botvinick & Cohen 1998) or the full-body illusion that can be created in most observers by synchronizing their actions and perceptions with those of an avatar seen in a VR (Virtual

Reality) presentation (see e.g., Blanke & Metzinger 2009; Blanke 2012; Botvinick & Cohen 1998; Lenggenhager et al. 2007). Only in such extreme cases does the error manifest itself in a complex relational case. We feel that we are someone or somewhere else and at the same time feel that we are not. It seems to take such extreme cases before we find a sizable illusion_{d+m} that deserves the name “illusion”.

In most cases, we can adjust perceptions once we notice that they are erroneous, be they ball trajectories or balancing properties. However, this adjustment process is painfully slow and may have to draw on early stages of perceptual and cognitive development. It does not take center stage, and some theoreticians would claim that the adjustment process converges on a veridical understanding of the world (Gibson 1979 calls this “attunement”). Others claim that many perceptions are useful precisely because they do not match or converge on the world (e.g., the multimodal user interface theory of perception, Hoffman 2010). The satisfying nature of private perception may not require a perfect solution of the UP in many cases, as long as the slips and falls remain limited to a tolerable number.

Acknowledgements

Markus Homberg, Cornelius Mülenz, and Mana Saadati helped collect and analyze the balance data; Daniel Oberfeld provided valuable input for the experimental designs; Elsa Krauß, Markus Landgraf, and Laura Längsfeld carried out the friction experiments.

References

- Blanke, O. (2012). Multisensory brain mechanisms of bodily self-consciousness. *Nature Reviews Neuroscience*, 13 (8), 556-571. [10.1038/nrn3292](https://doi.org/10.1038/nrn3292)
- Blanke, O. & Metzinger, T. (2009). Full-body illusions and minimal phenomenal selfhood. *Trends in Cognitive Sciences*, 13 (1), 7-13. [10.1016/j.tics.2008.10.003](https://doi.org/10.1016/j.tics.2008.10.003)
- Botvinick, M. & Cohen, J. (1998). Rubber hands 'feel' touch that eyes see. *Nature*, 391 (6669), 756-756. [10.1038/35784](https://doi.org/10.1038/35784)
- Chang, W. R. (1999). The effect of surface roughness on the measurement of slip resistance. *International Journal of Industrial Ergonomics*, 24 (3), 299-313. [10.1016/S0169-8141\(98\)00038-9](https://doi.org/10.1016/S0169-8141(98)00038-9)
- Chang, W. R., Grönqvist, R., Leclercq, S., Myung, R., Makkonen, L., Strandberg, L., Brungraber, R. J., Matthe, U. & Thorpe, S. C. (2001). The role of friction in the measurement of slipperiness, Part 1: Friction mechanisms and definition of test conditions. *Ergonomics*, 44 (13), 1217-1232. [10.1080/00140130110085574](https://doi.org/10.1080/00140130110085574)
- Clark, A. (2015). Embodied Prediction. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1-21). Frankfurt a. M., GER: MIND Group.
- Franz, V. H., Gegenfurtner, K. R., Bühlhoff, H. H. & Fahle, M. (2000). Grasping visual illusions: No evidence for a dissociation between perception and action. *Psychological Science*, 11 (1), 20-25. [10.1111/1467-9280.00209](https://doi.org/10.1111/1467-9280.00209)
- Frick, A., Rapp, A. F., Hug, S., Oláh, D. L. & Diggelmann, A. (2006). *Keine Haftung? Intuitives Wissen über Haftreibung bei Kindern und Erwachsenen*. Mainz, GER.
- Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton-Mifflin.
- Goldreich, D. (2007). A Bayesian perceptual model replicates the cutaneous rabbit and other tactile spatiotemporal illusions. *PLoS ONE*, 2 (3), e333-e333. [10.1371/journal.pone.0000333](https://doi.org/10.1371/journal.pone.0000333)
- Grierson, L. E. & Carnahan, H. (2006). Manual exploration and the perception of slipperiness. *Perception & Psychophysics*, 68 (7), 1070-1081. [10.3758/BF03193710](https://doi.org/10.3758/BF03193710)
- Grönqvist, R., Chang, W. R., Courtney, T. K., Leamon, T. B., Redfern, M. S. & Strandberg, L. (2001). Measurement of slipperiness: Fundamental concepts and definitions. *Ergonomics*, 44 (13), 1102-1117. [10.1080/00140130110085529](https://doi.org/10.1080/00140130110085529)
- Grüsser, O. J. (1983). Multimodal structure of the extrapersonal space. In A. Hein & M. Jeannerod (Eds.) *Spatially oriented behavior* (pp. 327-352). New York, NY: Springer.
- Hatfield, G. C. & Epstein, W. (1979). The sensory core and the medieval foundations of early modern perceptual theory. *Isis*, 70 (253), 363-384. [10.1086/352281](https://doi.org/10.1086/352281)
- Hecht, H. (2000). The failings of three event perception theories. *Journal for the Theory of Social Behaviour*, 30 (1), 1-25. [10.1111/1468-5914.00117](https://doi.org/10.1111/1468-5914.00117)
- (2013). Psychologische Anmerkungen zur Augentäuschung. In R. Steiner & C. Weissert (Eds.) *Lob der Illusion: Symposionsbeiträge*. München, GER: scaneg Verlag.
- Hecht, H. & Bertamini, M. (2000). Understanding projectile acceleration. *Journal of Experimental Psychology: Human Perception and Performance*, 26 (2), 730-746. [10.1037//0096-1523.26.2.730](https://doi.org/10.1037//0096-1523.26.2.730)
- Heller, M. A. (1982). Visual and tactual texture perception: Intersensory cooperation. *Perception & Psychophysics*, 31 (4), 339-344. [10.3758/BF03202657](https://doi.org/10.3758/BF03202657)
- Hermann, A. (1981). *Weltreich der Physik. Von Galilei bis Heisenberg*. Esslingen, GER: Bechtle Verlag.
- Hertwig, R. & Ortmann, A. (2005). The cognitive illusion controversy: A methodological debate in disguise that matters to economists. In R. Zwick & A. Rapoport (Eds.) *Experimental Business Research* (pp. 113-130). New York, NY: Springer.
- Hoffman, D. D. (2010). Sensory experiences as cryptic symbols of a multimodal user interface. *Activitas Nervosa Superior*, 52 (3-4), 95-104.
- Hohwy, J. (2015). The neural organ explains the mind. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1-22). Frankfurt a. M., GER: MIND Group.
- Joh, A. S., Adolph, K. E., Campbell, M. R. & Eppler, M. A. (2006). Why walkers slip: Shine is not a reliable cue for slippery ground. *Perception and Psychophysics*, 68 (3), 339-352. [10.3758/BF03193681](https://doi.org/10.3758/BF03193681)
- Joh, A. S., Adolph, K. E., Narayanan, P. J. & Dietz, V. A. (2007). Gauging possibilities for action based on friction. *Journal of Experimental Psychology: Human Perception and Performance*, 33 (5), 1145-1157. [10.1037/0096-1523.33.5.1145](https://doi.org/10.1037/0096-1523.33.5.1145)
- Kahneman, D., Slovic, P. & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge University Press.
- Kaiser, M. K., Proffitt, D. R., Whelan, S. & Hecht, H. (1992). Influence of animation on dynamical judgments. *Journal of Experimental Psychology: Human Perception and Performance*, 18 (3), 669-689. [10.1037/0096-1523.18.3.669](https://doi.org/10.1037/0096-1523.18.3.669)

- Kersten, D., Mamassian, P. & Yuille, A. (2004). Object perception as Bayesian Inference. *Annual Review of Psychology*, 55 (1), 271-304.
[10.1146/annurev.psych.55.090902.142005](https://doi.org/10.1146/annurev.psych.55.090902.142005)
- Knill, D. C. & Richards, W. (Eds.) (1996). *Perception as Bayesian inference*. New York, NY: Cambridge University Press.
- Lederman, S. J. & Abbott, S. G. (1981). Texture perception: Studies of intersensory organization using a discrepancy paradigm, and visual versus tactual psychophysics. *Journal of Experimental Psychology: Human Perception and Performance*, 7 (4), 902-915.
[10.1037/0096-1523.7.4.902](https://doi.org/10.1037/0096-1523.7.4.902)
- Lederman, S. J., Thorne, G. & Jones, B. (1986). Perception of texture by vision and touch: Multidimensionality and intersensory integration. *Journal of Experimental Psychology: Human Perception and Performance*, 12 (2), 169-180. [10.1037//0096-1523.12.2.169](https://doi.org/10.1037//0096-1523.12.2.169)
- Lederman, S. J. & Klatzky, R. L. (2009). Haptic perception: A tutorial. *Perception & Psychophysics*, 71 (7), 1439-1459. [10.3758/App.71.7.1439](https://doi.org/10.3758/App.71.7.1439)
- Lenggenhager, B., Tadi, T., Metzinger, T. & Blanke, O. (2007). Video ergo sum: Manipulating bodily self-consciousness. *Science*, 317 (5841), 1096-1099.
[10.1126/science.1143439](https://doi.org/10.1126/science.1143439)
- Martinez-Conde, S. & Macknik, S. L. (2010). Mind: The neuroscience of illusion. *Scientific American*, 20 (1), 4-6. [10.1038/nrn2473](https://doi.org/10.1038/nrn2473)
- McCloskey, M., Washburn, A. & Felch, L. (1983). Intuitive physics: The straight-down belief and its origin. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9 (4), 636-649.
[10.1037/0278-7393.9.4.636](https://doi.org/10.1037/0278-7393.9.4.636)
- Metzinger, T. (2003a). Phenomenal transparency and cognitive self-reference. *Phenomenology and the Cognitive Sciences*, 2, 353-393.
[10.1023/B:PHEN.0000007366.42918.eb](https://doi.org/10.1023/B:PHEN.0000007366.42918.eb)
- (2003b). *Being no one*. Cambridge, MA: MIT Press.
- Michaels, C. F. & Carello, C. (1981). *Direct perception*. Englewood Cliffs, NJ: Prentice-Hall.
- Milner, A. D. & Goodale, M. A. (2008). Two visual systems re-viewed. *Neuropsychologia*, 46 (3), 774-785.
[10.1016/j.neuropsychologia.2007.10.005](https://doi.org/10.1016/j.neuropsychologia.2007.10.005)
- Perreault, J. O. & Cao, C. G. (2006). Effects of vision and friction on haptic perception. *Hum Factors*, 48 (3), 574-586. [10.1518/001872006778606886](https://doi.org/10.1518/001872006778606886)
- Proffitt, D. R. & Linkenauger, S. A. (2013). Perception viewed as a phenotypic expression. In W. Prinz, M. Beisert & A. Herwig (Eds.) *Action science: Foundations of an emerging discipline* (pp. 171-197). Cambridge, MA: MIT Press.
- Reason, J. (1992). Cognitive underspecification. In B. J. Baars (Ed.) *Experimental slips and human error* (pp. 71-91). New York, NY: Springer US.
- Shepard, R. N. (1994). Perceptual-cognitive universals as reflections of the world. *Psychonomic Bulletin & Review*, 1 (1), 2-28. [10.3758/Bf03200759](https://doi.org/10.3758/Bf03200759)
- Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology*, 41, 1-19.
[10.1146/annurev.ps.41.020190.000245](https://doi.org/10.1146/annurev.ps.41.020190.000245)
- Thompson, P. & Wilson, J. (2012). Why do most faces look thinner upside down? *i-Perception* 3.10, 3, 765-774. [10.1068/I0554](https://doi.org/10.1068/I0554)
- Turvey, M. T. & Carello, C. (1995). Dynamic touch. In W. Epstein & S. Rogers (Eds.) *Handbook of perception and cognition (Vol. V, Perception of space and motion)* (pp. 401-490). San Diego, CA: Academic Press.
- Von Helmholtz, H. (1894). Über den Ursprung der richtigen Deutung unserer Sinneseindrücke. *Zeitschrift für Psychologie und Physiologie der Sinnesorgane*, 7, 81-91.
- Wason, P. C. & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content*. London, UK: Batsford.
- Wertheimer, W. (1912). Experimentelle Studien über das Sehen von Bewegung. *Zeitschrift für Psychologie*, 61, 161-265.
- Witt, J. K., Linkenauger, S. A., Bakdash, J. Z. & Proffitt, D. R. (2008). Putting to a bigger hole: Golf performance relates to perceived size. *Psychonomic Bulletin & Review*, 15 (3), 581-585. [10.3758/15.3.581](https://doi.org/10.3758/15.3.581)
- Wundt, W. (1898). Die geometrisch – optischen Täuschungen. *Akademie der sächsischen Wissenschaften Leipzig, Abhandlungen*, 24 (2), 54-178.

The Illusion of the Given and Its Role in Vision Research

A Commentary on Heiko Hecht

Axel Kohler

Illusions in vision and other modalities are captivating displays of the virtual nature of our subjective world. For this reason, illusions have been an important subject of scientific and artistic endeavors. In his target article, Heiko Hecht discusses the utility of the illusion concept and arrives at the negative conclusion that the traditional understanding of illusions as a discrepancy between world and perception is misguided. In his opinion, the more interesting and revealing cases are when the discrepancy is noticed and accompanies the perceptual state, or when, in the cognitive domain, the discrepancies become exceedingly large, but go unnoticed nonetheless. In this commentary, I argue that Hecht's criticism of the illusion concept is interesting and deserves further study. But at the current stage, I don't see that the model captures the essential features of illusory states. The processes on which Hecht focuses can be considered metacognitive appraisals of the respective sensory events, an interesting topic by itself. In the second part and as an overview, I review how research on the classical apparent-motion illusion has shaped our understanding of the neural underpinnings of motion perception and consciousness in general.

Keywords

Apparent motion | Bistability | Cognition | Illusion | Motion quartet | Multistability | Naïve realism | Perception | Phenomenal opacity | Phenomenal transparency | Sensation | Vection

Commentator

[Axel Kohler](#)

axelkohler@web.de

Universität Osnabrück
Osnabrück, Germany

Target Author

[Heiko Hecht](#)

hecht@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Illusions in science and culture

A main staple of research in cognitive science and especially vision science has been, and still is, the investigation of illusions. For one, it is just an amazing fact that although we think that our experience of the world is direct, we live by a subjective model of our environment. We feel that we perceive the world as it is, a naïve realism as we might call it, but we are just not aware that the world is only presented to us as a (re-)construction of our nervous system. In more philosophical terms, this fundamental property of our experience has been re-

ferred to as “phenomenal transparency” ([Metzinger 2003a](#)), the inability to recognize that our mental states are representations. This is probably the reason why we are baffled in cases when the subjective character of our perception becomes evident, although this rarely occurs under natural conditions.

At least in the context of our modern culture, many people will have had the experience that their train is leaving the station when in fact they have just watched the train on the opposite side of the platform taking off. This phe-

nomenon is termedvection, and everybody who has had this experience will remember the moment of insight when a cue destroys the illusion of self-motion and we realize that our train hasn't budged. A more historical example of illusions under natural conditions is the waterfall illusion—, a type of motion aftereffect. After looking at a waterfall or flowing water for a long time, static objects, e.g., the river bank or trees, seem to move in the direction opposite to the previously perceived water flow, probably due to adaptation effects in brain regions processing motion (Anstis et al. 1998). Early descriptions of the effect have been attributed to Aristotle (384–322 BCE) and Lucretius (99–55 BCE; Wade 1998). But apart from these few examples, it's rarely the case that the constructive nature of our perception is noticeable in everyday life.

Illusions have become a part of our popular culture and have had a strong impact on art. A whole art movement in painting, Op Art, is based on using known and discovering new visual illusions. It is a cultural version of vision research, presenting the fascinating nature of illusions to the public in aesthetically appealing ways. Illusions also feature prominently in the work of surrealist painter Salvador Dalí and other modern artists. For such artists, the medium presented a way of expressing the constructive nature of perception and signalled a departure from realism. For painters in general, knowledge about optics and the basis of visual perception has always been important for guiding the construction process of paintings and the refinement of techniques in order to achieve certain effects in the eye of the beholder. The entwinement of science and art is scrutinized in recent work looking at the interaction between fields (Zeki 1999). Two other forms of art that were more or less invented in close interaction with science are photography and film-making. The very basis of TV and movie presentations is rooted in the fact that we are able to fuse a rapid sequence of static images to construct a natural impression of moving objects. TV displays, projectors, and computer screens work with a certain refresh rate at which subsequent images are presented; the rate can be as low as

24 Hz in cinematography. The basic phenomenon that allows us to create a natural perceptual flow from flickering images is referred to as apparent motion, a type of illusory motion.

Because of the fascination with illusions and its influence on culture, illusions have been guiding research on visual perception for a long time—and continue to do so. But this is not the only reason for the utilization of illusions in science. Illusions are a powerful tool for understanding mechanisms of sensory processing in the brain that are unexpected or counterintuitive. Many motion illusions where motion can be seen in static displays (often seen in the entertainment sections of magazines) depend on a specific configuration of color values in directly abutting picture elements. These configurations of picture elements are repeated and cover the entire display, in sum creating a striking motion impression. Psychophysical experiments showed that the key to the illusion is the configuration of neighboring elements, whose effects cannot be predicted by current models of visual processing. Additional neurophysiological measurements in the same study demonstrated that different picture elements were processed with different latencies in certain areas of the visual cortex, mimicking a motion signal (Conway et al. 2005). This suggested a neural explanation for the occurrence of the illusion and led to a revision of existing models of motion selectivity.

Another driving force for the use of illusions in research was a resurgence of interest in understanding conscious perception. At the beginning of the 1990s, Francis Crick and Christof Koch started to publish a sequence of conceptual papers advocating the investigation of consciousness with empirical, and especially neuroscientific methods (Crick & Koch 1990, 1995, 1998). Since then the number of papers on consciousness has grown steadily in the domain of cognitive neuroscience. Certain visual illusions lend themselves specifically to investigating the nature of conscious processing. Some of the most prominent paradigms display the characteristic of bistability or multistability: When presented to observers, conscious perception alternates between two (bistability) or multiple (multistability) interpretations although the

physical characteristics of the display do not change. Rubin's face-vase illusion and the Necker Cube are just the most prominent among a multitude of examples for multistability (Kim & Blake 2005). The promise of using multistability is that it allows for disentangling the neural representation of the physical stimulus characteristics from the processes giving rise to conscious perception. The logic of the approach is that changes in neural activity accompanying switches in subjective experience during constant physical stimulation provide a guide to understanding the neural underpinnings of consciousness.

2 Hecht's criticism of the illusion concept

In his target article "Beyond illusions: On the limitations of perceiving relational properties," Heiko Hecht (this collection) begins with a discussion of the traditional concept of illusion and how it has been employed in the context of research on vision. In its most basic sense, an illusion refers to a difference between our representation of a given scene and its actual physical properties. In an interesting take on the utility of illusions in research, Hecht suggests that the mere discrepancy between our perception and the real world—what he calls *illusion_d* ("d" for "discrepancy")—is less useful than one might think. In simple terms, our perception is off to some degree in many cases. But still, on the other hand it is amazing how on-target it is most of the time: it is sufficiently accurate for an effective interaction with the world. For Hecht, the term "illusion" should be reserved for situations when discrepancies (*illusion_d*) are manifest, i.e., when the error is part of the experience and we become aware of it. This is termed *illusion_m* ("m" for "manifest") and is supposed to be the more interesting case. The moment of insight for the train-ride illusion described above might be a good example. Interpreting relative motion between trains as self-motion is often an adequate interpretation, but the error is manifested in a striking fashion experientially when we spot a part of the platform that indicates unmistakably that we are still in the same place.

In addition to the distinction between *illusion_d* and *illusion_m*, Hecht is concerned with cognitive illusions in comparison to the well-known perceptual illusions. His interesting observation is that when we move away from perception, the discrepancies between the real world and our judgments become even larger, sometimes to an absurd level. Humans are notoriously bad at everyday physics. Hecht mentions that we see nothing wrong with fabricated scenes that glaringly contradict Newtonian physics, and even our spontaneous actions reveal the same degree of error. Nevertheless, they are hardly ever noticed, i.e., *illusion_d* rarely becomes *illusion_m* in the cognitive domain. That this is especially the case for relational properties Hecht demonstrates with a series of his own experiments on physics judgments by university students. Even participants that should at least have some theoretical knowledge about the laws governing the real world (physics students) are surprisingly bad at finding the right answers to quizzes on balancing beams made of different materials with different weight distributions (Experiment 1) and on the slipperiness of surfaces (Experiment 2). In these examples, the students' judgments are in stark contrast to the actual, real-world outcomes, which were also empirically tested in addition to deriving predictions from the laws of physics. So even though the paradigms were chosen to be experientially accessible and ecologically relevant, it seems that our cognitive system does not care about correctness or even rough approximations that would point it in the right direction. Even the mere ordering of solutions without providing quantitative details is seldom correct.

To summarize, Hecht suggests that the small deviations of our perceptual representations are no match for the sometimes extreme discrepancies found in the cognitive domain. *Illusion_d* is the norm rather than the interesting exception in sensory processing because—at least in vision—the full three-dimensional representation of the world has to be derived from a limited array of two-dimensional information on the retinae. Hecht (this collection) refers to this as the "underspecification problem." For an efficient solution to the underspecification prob-

lem, the system employs a range of assumptions and constraints on the makeup of the world to guide the reconstruction process. For Hecht, perception is therefore always fraught with cognitive elements. This is even more so when discrepancy is detected; *illusion_d* becomes *illusion_m*. Then, cognitive judgments are involved, and an explicit comparison process is initiated that allows us to capture the discrepancy and which makes it experientially available.

3 The role of illusions in vision research

Hecht provides compelling evidence for the error-prone nature of everyday judgments, especially when it comes to relational properties. His observation of an antagonism between the size of discrepancies and their detectability is interesting. Moving from the perceptual to the cognitive domain, the size of discrepancies increases, but at the same time we are less likely to notice those errors. But there are a few points of dissent I would like to discuss in what follows. (1) The discussion of the cognitive nature of perception is long-standing and won't be solved in the near future, especially because the term "cognition" is notoriously imprecise. Nevertheless, I am not convinced that the cognitive aspect that is supposed to be part of perceptual as well as cognitive illusions in Hecht's view is a necessary ingredient for a proper concept of illusion. (2) Hecht's arguments are a welcome incentive to reflect upon the concept of illusion and its role for research. Although he does not negate the role of perceptual illusions for vision research, he is rather critical concerning the utility of traditional illusion research, especially with respect to the underspecification problem. Drawing on the vast body of research on apparent motion, I would like to provide an example of a positive research program that has accumulated valuable insights into the mechanisms underlying visual motion processing. This is not necessarily in contradiction to Hecht's stance. The focus of research on illusions has focused more on the neural mechanisms of visual processing and specifically on the neural correlates of conscious perception. In this sense, the research lines can be seen as complementary.

Nevertheless, I would argue in conclusion that the term illusion is well anchored in the perceptual domain and plays an important guiding role for research on visual processing.

There is a long tradition in vision research of considering the influence of cognition on perceptual processes. The basis for the early investigations on vision and, more generally, on sensory processing in the 19th century and early 20th century was the distinction between sensation and perception. One of [Helmholtz's \(1863\)](#) definitions captures the main line of thought:

Empfindungen nennen wir die Eindrücke auf unsere Sinne, insofern sie uns als Zustände unseres Körpers (speciell unserer Nervenapparate) zum Bewusstsein kommen; Wahrnehmungen insofern wir uns aus ihnen die Vorstellung äusserer Objecte bilden.¹

The definition can be seen as a continuation of a philosophical tradition that has the intention of separating pure states of sensory reception from the more cognitive aspects concerned with the reconstruction of the outer world. Already at this time, different authors were aware of the fact that these definitions did not draw a clear dividing line between different types of sensory states. For example, [Sigmund Exner \(1875\)](#) refers to Helmholtz's definition and points to several examples for which the distinction becomes muddled. His observant conclusion is that the philosophical concepts do not fare well in the field of brain physiology and that contradictions have to be resolved in future models of sensory processing ([Exner 1875](#), p. 159). So despite its initial allure, the distinction between sensation and perception produced more problems than solutions.

An interesting recent model of the interaction between perception and cognition has been proposed by [Vetter & Newen \(2014\)](#). They review the current empirical literature on cognitive penetration of perceptual processing and

¹ English:

"We call the impressions on our senses sensations, insofar as we become aware of them as states of our body (especially of our nervous system); we call them perceptions insofar as we create representations of external objects." [My translation]

find compelling evidence that cognitive penetration of perception is ubiquitous. They distinguish four stages of processing in the sensory (visual) hierarchy: (1) basic feature detection, (2) percept estimation, (3) learned visual patterns, and (4) semantic world knowledge. According to their account, almost all possible interactions between processing levels occur under normal conditions and top-down connections can be considered forms of cognitive penetration. They argue that it's not a question of whether cognition influences perception, but rather of what type of interaction takes place in any given case. They advocate a move away from the general conceptual question of the cognition-perception relationship towards an empirically-based consideration of the interactions between different levels of the processing hierarchy.

Importantly, none of the stages characterized by Vetter & Newen (2014) capture the cognitive component Hecht has in mind. The realization that there is a discrepancy between percept and the real world is not something involved in the construction of the perceptual content itself. It seems that this it is more along the lines of a metacognitive appraisal of the current situation. With reference to Metzinger's (2003b) concept of phenomenal transparency (a naïve-realistic stance towards the perceived world) referred to at the beginning of the commentary, it is now the complementary feature of phenomenal opacity—a situation in which the representational character of experience becomes available to the subject—that might play a role here. Metzinger (2003b) refers to cases of lucid dreaming and drug-induced hallucinations as prime examples of phenomenal opacity. Interestingly, it is not sufficient for him that we have accompanying reflexive thoughts on the nature of perceptual representations (the “philosopher's stance”, as one could say), but we must also be attentively engaged with the perceptual content and recognize the illusory nature of the process. Therefore, it seems to be the case that neither the views of Vetter & Newen nor Metzinger's concept of phenomenal opacity seem to capture the cognitive component Hecht has in mind. But in my view, such models of cognitive penet-

ration are much more intimately linked with the illusion concept, because they provide an understanding of how the very nature of the experience is modulated by cognitive processes. Hecht's model doesn't seem to capture that aspect, since it functions more as a cognitive commentary on the impenetrable perceptual process. It is unclear why this metacognitive appraisal should be considered a hallmark of illusory experiences.

When Hecht argues for abandoning the term “illusion” in the perceptual domain, he also refers to Wertheimer's classical work on apparent motion (1912) and contends that the Gestalt psychologists “avoided the term illusion” (Hecht this collection). It is true that, for example, Wertheimer (1912, pp. 167–168) himself mentions in a footnote that “illusion” should not be used to refer to a discrepancy relative to the physical world because his main concern is with mental states. (The German word in the original paper is “Täuschung,” which is indeed best translated as “illusion” in this context.) Nevertheless, the passage is not very clear on the reasons for rejecting the reference to discrepancy. Again, it seems that the distinction between sensation and perception (see above) is lingering in the background. Even assuming a correct sensory reception (sensation) of the apparent-motion inducers, something is added that goes beyond the raw sensory data. In a later section of the paper (Wertheimer 1912, p. 228), this becomes clearer when Wertheimer analyzes another possible meaning of “Täuschung,” i.e., failure of judgment (German: “Urteilstäuschung”). It is important for him that apparent motion is not a result of cognitive processes, of inferences of the type: “If an object was there just before and now is over here, it must have moved between the points.” He is convinced of the perceptual nature of the phenomenon and rejects the idea that cognition plays an important role. Again, there is some ambiguity with respect to the usage of the term “illusion” here. This being said, throughout the article Wertheimer uses the noun-form “Täuschung” thirty-five times and also refers to other motion illusions that were already as well known at the time as “Täuschung.” On my

reading, his main intention was to prove that apparent motion is the result of a low-level perceptual process and that it is indeed illusory in nature.

Wertheimer's 1912 paper, with its detailed psychophysical investigation of the apparent-motion phenomenon is commonly considered to be the founding event of the Gestalt movement (cf. Sekuler 1996; Steinman et al. 2000), although this might not be the complete picture (Wertheimer 2014). We have just passed the centenary of Wertheimer's seminal article, but still there is much work to be done to provide a complete picture of the processes involved in apparent-motion perception at behavioral, computational, and neurophysiological levels of description. In my view, apparent motion is a paradigmatic case of an *illusion_d* that has fertilized the understanding of motion processing and continues to do so. Given the roughly one hundred years of research on apparent motion, it is worthwhile to take stock (briefly) and see where investigations associated with this paradigm have taken us.

Psychophysical investigations of apparent motion are too numerous to review extensively here. Early studies focused on describing the basic features of the phenomenon. Korte's laws (1915) are still part of textbook knowledge in vision research; he described the influence of different stimulus characteristics (stimulus strength, spatial and temporal separation etc.) on the quality of apparent-motion perception. New varieties of apparent motion were described in the following, one of the most important ones being the motion quartet (Neuhaus 1930; von Schiller 1933; see video: <http://www.open-mind.net/videomaterials/kohler-motion-quartet>). This is a bistable version of apparent motion, where two frames with diagonally opposing dots at the corners of a virtual rectangle are flashed in alternation. The identical stimulus sequence can be interpreted as being in vertical or horizontal motion. During longer presentations of the unchanging stimulus, conscious perception will spontaneously switch between the possible alternatives. It is therefore an important example of a multistable display, which allows

various interpretations with the same physical input. Early on, it was noticed that the integration of motion inducers in the motion quartet processed within brain hemispheres is facilitated relative to integration between hemispheres (Gengerelli 1948), a fact we will come back to later on. After a relative hiatus in the 50s and 60s, apparent motion again took center stage in the 70s. It was the basis for the work of Paul Kolars (1972) on configuration effects and for the first investigation of computational principles of motion perception by Shimon Ullman (1979). At the same time, distinctions between different types of apparent motion were introduced (Anstis 1980; Braddick 1974, 1980), later culminating in the three-layered hierarchical system of motion types proposed by Lu & Sperling (1995, 2001).

Currently, in all domains (psychophysical, computational, neurophysiological) there are ongoing research endeavors cross-fertilizing each other in the search for mechanisms underlying illusory perception of motion. After the turn of the millennium, the broad availability of brain-imaging methods spurred the investigation of the neural mechanisms underlying apparent-motion perception. By and large, the same areas that process real motion are involved in the (Muckli et al. 2002; Sterzer et al. 2003; Sterzer et al. 2002; Sterzer & Kleinschmidt 2005), supporting the assumption that results from studies on apparent motion can be transferred to other types of motion processing. Another interesting result from studies using functional magnetic resonance imaging was that traces of the virtual apparent-motion path, the illusory motion between inducers, can already be seen in the primary visual cortex, the earliest stage of visual cortical processing (Larsen et al. 2006; Muckli et al. 2005). This effect is probably mediated through feedback connections from higher areas (Sterzer et al. 2006), explaining the fact that normal visual functioning is disturbed on the path of apparent motion (Yantis & Nakama 1998) and also supporting Wertheimer's (1912) original claim that apparent motion is a perceptual phenomenon that does not depend on cognitive inferences. Animal studies are starting to elucidate the more fine-grained neural

mechanisms subserving apparent-motion processing. Neurophysiological investigations in the animal model demonstrated complex wave patterns of interactions between several cortical areas during the perception of apparent motion (Ahmed et al. 2008). This work also inspired a formal model of these interactions elucidating the computational principles underlying the representation of apparent motion in the brain (Deco & Roland 2010).

In my own recent research, I have specifically looked at interindividual differences in the perception of apparent motion and its anatomical basis. As mentioned above, for the bistable motion quartet there is a difference between perceiving apparent motion in the vertical and horizontal direction. Observers show a bias towards perceiving vertical motion when they fixate on the middle of the motion quartet (Chaudhuri & Glaser 1991). A possible explanation for this is that due to the way the visual field is represented in the visual cortex, vertical motion only requires integration within brain hemispheres, but horizontal motion depends on integration between hemispheres. In fact, we could demonstrate that the individual bias of observers of vertical motion could be partly predicted by the quality of the neural connections between brain halves, suggesting that interhemispheric integration is a relevant factor (Geng et al. 2011).

The very short summary of research on apparent motion demonstrates the various insights this simple paradigm has inspired over the course of the last century and beyond. It led to a detailed description of the involved brain areas, including interindividual differences, and to processing models being developed on the computational and neurophysiological level. As mentioned in the introductory section, one main concern in vision research associated with illusions is the interest in conscious perception and the property of multistability. Both aspects are also dominant in the apparent-motion field. The current state of research is just the starting point for investigations towards a deeper understanding of the exact mechanisms. Often, the results are still descriptive and qualitative in nature and don't allow for very specific predic-

tions with respect to the involved neural machinery and dynamics. Yet the research line is promising and has the potential to lead to broadly applicable results. This might even be the case for the underspecification problem, the problem of reconstructing a full-fledged 3D world from a limited 2D input—one of Hecht's main concerns. Multistability can be seen as one paradigm case in which the nervous system has to resolve ambiguity. For the Necker Cube, the motion quartet, and other multistable displays, the brain settles into a solution for a perceptual problem by resolving competition among alternatives. Therefore, research on multistability might help to elucidate the core mechanisms that give rise to the definitive subjective interpretations with which we represent the world.

4 Conclusion

In conclusion, Hecht's distinction between *illusion_d* and *illusion_m* and his criticism of the naïve illusion concept in vision research is interesting. When we become aware of illusions, when we suddenly recognize the virtual character of our subjective world, certain metacognitive processes are initiated that are a worthwhile subject matter for further investigation. In some sense they become part of the experience, and an important question is whether and how the two aspects of the experience interact. Nevertheless, Hecht also agrees that perceptual representations are relatively immune to top-down control, i.e., even in the rare cases in which the illusory character becomes manifest, the perceptual processes are mostly modular and impenetrable in nature. Therefore, the question of illusory representation can be tackled independently of the question of metacognitive awareness, and continues to be an important guide for research on visual processing. Apart from looking at the more conceptual question of the level at which the term “illusion” should be applied, which is moot to some degree, I have tried to provide examples of relevant *illusion_d* research that has made progress on the question of how the brain processes visual information. Even for the underspecification problem, there is opportunity for valuable insight, which hasn't been exploited to full potential yet in current research.

References

- Ahmed, B., Hanazawa, A., Undeman, C., Eriksson, D., Valentiniene, S. & Roland, P. E. (2008). Cortical dynamics subserving visual apparent motion. *Cerebral Cortex*, 18 (12), 2796-2810. [10.1093/cercor/bhn038](https://doi.org/10.1093/cercor/bhn038)
- Anstis, S. M. (1980). The perception of apparent movement. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences*, 290 (1038), 153-168. [10.1098/rstb.1980.0088](https://doi.org/10.1098/rstb.1980.0088)
- Anstis, S. M., Verstraten, F. A. J. & Mather, G. (1998). The motion aftereffect. *Trends in Cognitive Sciences*, 2 (3), 111-117. [10.1016/S1364-6613\(98\)01142-5](https://doi.org/10.1016/S1364-6613(98)01142-5)
- Braddick, O. (1974). A short-range process in apparent motion. *Vision Research*, 14 (7), 519-527. [10.1016/0042-6989\(74\)90041-8](https://doi.org/10.1016/0042-6989(74)90041-8)
- (1980). Low-level and high-level processes in apparent motion. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences*, 290 (1038), 137-151. [10.1098/rstb.1980.0087](https://doi.org/10.1098/rstb.1980.0087)
- Chaudhuri, A. & Glaser, D. A. (1991). Metastable motion anisotropy. *Visual Neuroscience*, 7 (5), 397-407. [10.1017/S0952523800009706](https://doi.org/10.1017/S0952523800009706)
- Conway, B. R., Kitaoka, A., Yazdanbakhsh, A., Pack, C. C. & Livingstone, M. S. (2005). Neural basis for a powerful static motion illusion. *Journal of Neuroscience*, 25 (23), 5651-5656. [10.1523/JNEUROSCI.1084-05.2005](https://doi.org/10.1523/JNEUROSCI.1084-05.2005)
- Crick, F. & Koch, C. (1990). Some reflections on visual awareness. *Cold Spring Harbor Symposia on Quantitative Biology*, 55, 953-962. [10.1101/SQB.1990.055.01.089](https://doi.org/10.1101/SQB.1990.055.01.089)
- (1995). Are we aware of neural activity in primary visual cortex? *Nature*, 375 (6527), 121-123. [10.1038/375121a0](https://doi.org/10.1038/375121a0)
- (1998). Consciousness and neuroscience. *Cerebral Cortex*, 8 (2), 97-107. [10.1093/cercor/8.2.97](https://doi.org/10.1093/cercor/8.2.97)
- Deco, G. & Roland, P. (2010). The role of multi-area interactions for the computation of apparent motion. *NeuroImage*, 51 (3), 1018-1026. [10.1016/j.neuroimage.2010.03.032](https://doi.org/10.1016/j.neuroimage.2010.03.032)
- Exner, S. (1875). Über das Sehen von Bewegungen und die Theorie des zusammengesetzten Auges. *Sitzungsberichte der Akademie der Wissenschaften Wien (Mathematisch-Naturwissenschaftliche Klasse)*, 72 (3), 156-190.
- Gengerelli, J. A. (1948). Apparent movement in relation to homonymous and heteronymous stimulation of the cerebral hemispheres. *Journal of Experimental Psychology*, 38 (5), 592-599. [10.1037/h0062438](https://doi.org/10.1037/h0062438)
- Genç, E., Bergmann, J., Singer, W. & Kohler, A. (2011). Interhemispheric connections shape subjective experience of bistable motion. *Current Biology*, 21 (17), 1494-1499. [10.1016/j.cub.2011.08.003](https://doi.org/10.1016/j.cub.2011.08.003)
- Hecht, H. (2015). Beyond illusions: On the limitations of perceiving relational properties. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Kim, C. Y. & Blake, R. (2005). Psychophysical magic: Rendering the visible “invisible.”. *Trends in Cognitive Sciences*, 9 (8), 381-388. [10.1016/J.Tics.2005.06.012](https://doi.org/10.1016/J.Tics.2005.06.012)
- Kolers, P. A. (1972). *Aspects of motion perception*. New York, NY: Pergamon Press.
- Korte, A. (1915). Kinematoskopische Untersuchungen. *Zeitschrift für Psychologie*, 72, 194-296.
- Larsen, A., Madsen, K. H., Lund, T. E. & Bundesen, C. (2006). Images of illusory motion in primary visual cortex. *Journal of Cognitive Neuroscience*, 18 (7), 1174-1180. [10.1162/jocn.2006.18.7.1174](https://doi.org/10.1162/jocn.2006.18.7.1174)
- Lu, Z. L. & Sperling, G. (1995). Attention-generated apparent motion. *Nature*, 377 (6546), 237-239. [10.1038/377237a0](https://doi.org/10.1038/377237a0)
- (2001). Three-systems theory of human visual motion perception: Review and update. *Journal of the Optical Society of America A, Optics, Image Science, and Vision*, 18 (9), 2331-2370. [10.1364/JOSAA.18.002331](https://doi.org/10.1364/JOSAA.18.002331)
- Metzinger, T. (2003a). *Being no one: The self-model theory of subjectivity*. Cambridge, MA: MIT Press.
- (2003b). Phenomenal transparency and cognitive self-reference. *Phenomenology and the Cognitive Sciences*, 2 (4), 353-393. [10.1023/B:PHEN.0000007366.42918.eb](https://doi.org/10.1023/B:PHEN.0000007366.42918.eb)
- Muckli, L., Kriegeskorte, N., Lanfermann, H., Zanella, F. E., Singer, W. & Goebel, R. (2002). Apparent motion: Event-related functional magnetic resonance imaging of perceptual switches and states. *Journal of Neuroscience*, 22 (9), RC219.
- Muckli, L., Kohler, A., Kriegeskorte, N. & Singer, W. (2005). Primary visual cortex activity along the apparent-motion trace reflects illusory perception. *PLoS Biology*, 3 (8), e265. [10.1371/journal.pbio.0030265](https://doi.org/10.1371/journal.pbio.0030265)
- Neuhaus, W. (1930). Experimentelle Untersuchung der Scheinbewegung. *Archiv für die gesamte Psychologie*, 75, 315-458.
- Sekuler, R. (1996). Motion perception: A modern view of Wertheimer's 1912 monograph. *Perception*, 25 (10), 1243-1258. [10.1068/p251243](https://doi.org/10.1068/p251243)
- Steinman, R. M., Pizlo, Z. & Pizlo, F. J. (2000). Phi is

- not beta, and why Wertheimer's discovery launched the Gestalt revolution. *Vision Research*, 40 (17), 2257-2264. [10.1016/S0042-6989\(00\)00086-9](https://doi.org/10.1016/S0042-6989(00)00086-9)
- Sterzer, P., Russ, M. O., Preibisch, C. & Kleinschmidt, A. (2002). Neural correlates of spontaneous direction reversals in ambiguous apparent visual motion. *NeuroImage*, 15 (4), 908-916. [10.1006/nimg.2001.1030](https://doi.org/10.1006/nimg.2001.1030)
- Sterzer, P., Eger, E. & Kleinschmidt, A. (2003). Responses of extrastriate cortex to switching perception of ambiguous visual motion stimuli. *Neuroreport*, 14 (18), 2337-2341. [10.1097/01.wnr.0000102554.45279.a3](https://doi.org/10.1097/01.wnr.0000102554.45279.a3)
- Sterzer, P., Haynes, J.-D. & Rees, G. (2006). Primary visual cortex activation on the path of apparent motion is mediated by feedback from hMT+/V5. *NeuroImage*, 32 (3), 1308-1316. [10.1016/j.neuroimage.2006.05.029](https://doi.org/10.1016/j.neuroimage.2006.05.029)
- Sterzer, P. & Kleinschmidt, A. (2005). A neural signature of colour and luminance correspondence in bistable apparent motion. *European Journal of Neuroscience*, 21 (11), 3097-3106. [10.1111/j.1460-9568.2005.04133.x](https://doi.org/10.1111/j.1460-9568.2005.04133.x)
- Ullman, S. (1979). *The interpretation of visual motion*. Cambridge, MA: MIT Press.
- Vetter, P. & Newen, A. (2014). Varieties of cognitive penetration in visual perception. *Consciousness and Cognition*, 27, 62-67. [10.1016/j.concog.2014.04.007](https://doi.org/10.1016/j.concog.2014.04.007)
- von Helmholtz, H. (1863). *Die Lehre von den Tonempfindungen, als physiologische Grundlage für die Theorie der Musik*. Braunschweig, GER: Vieweg.
- von Schiller, P. (1933). Stroboskopische Alternativversuche. *Psychologische Forschung*, 17 (1), 179-214. [10.1007/BF02411959](https://doi.org/10.1007/BF02411959)
- Wade, N. J. (1998). *A natural history of vision*. Cambridge, MA: MIT Press.
- Wertheimer, M. (1912). Experimentelle Studien über das Sehen von Bewegung. *Zeitschrift für Psychologie*, 61, 161-265.
- (2014). Music, thinking, perceived motion: The emergence of Gestalt theory. *History of Psychology*, 17 (2), 131-133. [10.1037/a0035765](https://doi.org/10.1037/a0035765)
- Yantis, S. & Nakama, T. (1998). Visual interactions in the path of apparent motion. *Nature Neuroscience*, 1 (6), 508-512. [10.1038/2226](https://doi.org/10.1038/2226)
- Zeki, S. (1999). *Inner vision: An exploration of art and the brain*. Oxford, UK: Oxford University Press.

Manifest Illusions

A Reply to Axel Kohler

Heiko Hecht

The notion of illusion as a discrepancy between physical stimulus and percept (here referred to as illusion_d , as long as merely this “error” is meant) is unable to capture the four very different cases in which illusions can arise. The observer may or may not be aware of the discrepancy, and its magnitude may be large or small. I argue that the special case of small error paired with awareness deserves special attention. Only in this case does the observer readily see the illusion, since it becomes manifest (referred to as illusion_m). Illusion_m is a meaningful category even in cases where illusion_d cannot be determined. Illusions_m of apparent motion and illusions of intuitive physics are solicited.

Keywords

Apparent motion | Illusion | Illusion_m | Intuitive physics | Manifest illusions | Relational properties | Underspecification problem

Author

Heiko Hecht

hecht@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Commentator

Axel Kohler

axelkohler@web.de

Universität Osnabrück
Osnabrück, Germany

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 The concept of illusion

Axel Kohler points out that illusions understood as discrepancy between physical stimulus and percept (illusion_d) have inspired progress in the history of experimental psychology. At first glance, this seems to be rather obvious. However, to define a discrepancy, one must have two comparable measures of the same thing. But this is often not the case. Take a given lamp that looks very dim to us during the day but blindingly bright at night. How bright is the stimulus really? We are unable to determine which of the two cases is more illusory_d. The perceiver does not normally notice the illusion_d . Apparent motion, in contrast, which has been a

very influential paradigm, is more than mere illusion_d . By differentiating illusions into illusion_d and illusion_m , I am able to point out a strange inconsistency between the amount of error contained in an illusion and the perceptual conspicuity of this error. I argue that there are four varieties of discrepancy between physical stimulus and the related percept (illusion_d). They can be grouped by the size of the discrepancy and the degree of awareness (see Figure 1). First, there are more or less subtle discrepancies that are ubiquitous and go unnoticed most of the time. In rare occasions, and usually triggered by a revealing piece of contradiction, they are no-

ticed (illusion_m). The second variety consists of very large discrepancies, such as found in many intuitive physics examples. For instance, a water surface may look fine even if it extends impossibly at a large angle from the horizontal. For instance, when asked to draw the surface level that water assumes in a tilted beaker, observers err as if they did not know that water remains parallel to the ground. And the more expert they become at avoiding spills, the larger the error becomes. Experienced bartenders produce the largest errors (see [Hecht & Proffitt 1995](#)). The perception of relational properties discussed in the target article falls into this category. Here the perceptual error can be enormous and still go unnoticed. Typically, we need to consult physics books and learn about a physical stimulus before we are convinced that our perception is erroneous. When conceiving of illusion as mere illusion_d, we fail to honor the special case of illusion_m. Illusions_d are ubiquitous. As a matter of fact, the core discipline of sensory psychology—psychophysics—can be thought of as the formal description of how a physical stimulus differs from its percept. It does so all the time. Illusions_m are a special case. They may warn the organism about where adjustments to the perceptual system are necessary in order to avoid potentially dangerous misjudgments. Or they may just be occasions where the perceptual system fails to suppress the perceptual process that has lost out in the competition to resolve the underspecification problem.

2 Apparent motion (AM)

I thank Axel Kohler for bringing up AM (apparent motion) as an example of how seminal an illusion can be for research. I do concur that it continues to be a fascinating phenomenon. However, I believe that AM did not fascinate [Wertheimer \(1912\)](#) because it is an illusion_d, but rather because it is predominantly an illusion_m. Note that the timing has to be just so (i.e., a particular combination of on-times and ISI, inter-stimuli-intervals) in order to perceive what he called phi-motion: perfectly smooth motion practically indistinguishable from real motion.

Most of his experiments and demonstrations have in fact worked with suboptimal cases in which the perceived motion is bumpy or faint. In all these other cases of AM, the illusory nature of the percept becomes manifest. The bistable quartet is another beautiful case of an illusion_m. The mere fact that the percept can flip at will shows the illusion_m to be manifest.

illusions

		Awareness	
		YES	NO
Magnitude of the discrepancy (stimulus–percept)	SMALL	illusion _m	opaque illusion _d
	LARGE	insight	intuitive physics psychophysics

Figure 1: Varieties of illusions.

As an aside, the Gestalt laws can be understood as an attempt to describe how the percept emerges from the given physical stimulus. But note that while the percept is always different from the physical stimulus, it should not be thought of as illusory just because it is the outcome of a Gestalt process. When I said that Gestalt psychologists have “avoided the term illusion” I was not expecting anyone to count the occurrences of the term in Wertheimer’s 1912 paper. He did use the term. I stand corrected. Note, however, that he put the term “Täuschung” in quotation marks the first time he used it, well aware that the phenomenal experience of motion is what makes the Gestalt, regardless of how it relates to the physical stimulus.

Another revealing aspect of AM is its power to reveal the extent to which world-know-

ledge is factored into our perception, unconsciously and the more so the less well-defined the stimulus. Let us consider a classic AM-display in which two rectangles at two locations and at different orientations are shown in alternation. Whenever the ISI is short (say 100ms), we see one rectangle moving on a straight path and changing its orientation concurrently. If, however, the ISI is lengthened (to, say, 500ms), then the AM-path curves (see [McBeath & Shepard 1989](#); [Hecht & Proffitt 1991](#)). The phenomenal quality of this motion is rather ephemeral. We immediately see that the motion is not distinct but fraught with uncertainty. When choosing intermediate ISI, and forcing observers to make up their minds, some observers will see the rectangle curve and others will see it move along a straight path. And when the display remains unchanged but the area between the rectangles is shaded, then the rectangle appears to move along the shaded path. Thus, one can direct the motion of the rectangle along almost arbitrary paths (e.g., [Shepard & Zare 1983](#)). Such demonstrations reveal that the very notion of error or discrepancy between physical stimulus and percept becomes shaky. It seems rather arbitrary whether the researcher considers only the rectangles to be the relevant stimulus or also considers the background to be part of the stimulus. In these AM displays, the visual system appears to make sense of the entire display, not just the two moving rectangles.

3 The case for illusion_m

Such resolution of the underspecification problem can even annihilate an existing illusion_d. Consider the sophisticated AM display we encounter when going to the movies. And let us take the old-fashioned kind, where the projection screen is black most of the time, only interrupted 24 times a second by a very brief flash of a stationary picture. Smooth motion is perceived. Here, the observer is typically unaware of the illusion_d, but what is perceived is actually closer to the original scene than to the movie that was made from it. We might even entertain the idea that there is no illusion_d, since the percept is very close to the original scene that was

filmed. Now, calling apparent motion illusory_d when dealing with artificial or computer-generated stimuli, but veridical when dealing with a movie, does not seem to make much sense. This is because, in a very deep sense, the visual system has no way of distinguishing between actual motion and snapshot motion. The hardware we use to detect motion is built such that it is unable to differentiate between the two. Basically, the detector for motion is designed such that successive excitations of the receptive fields of two motion-sensitive neurons lead to the impression of motion. These Reichardt/Hassenstein detectors ([Hassenstein & Reichardt 1956](#)) are discrete; they cannot tell the difference between continuous and stroboscopic motion (see e.g., [Hecht 2006](#)). Note that this holds for phi-motion but falls apart when ISI or duty cycle are changed.

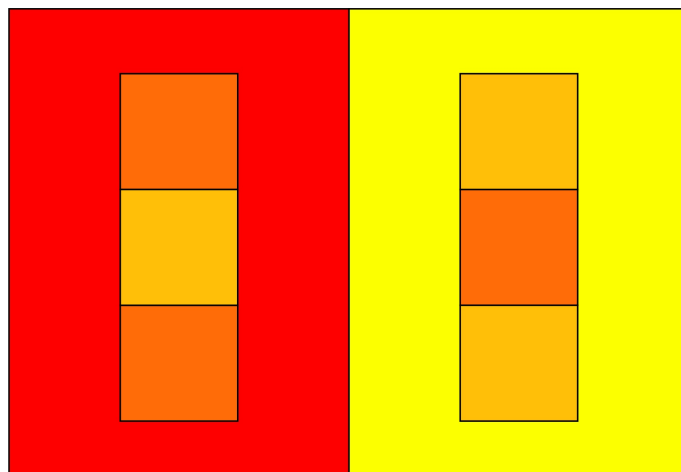


Figure 2: Simultaneous color contrast: The orange and the yellow squares are of the same respective color in the panel on the left and on the right.

Let us now look at an example from the color domain to further challenge the notion of illusion_d. The phenomenon of color constancy lets us perceive the same color even if the ambient lighting changes dramatically. We see an object as blue regardless of whether the room is lit by a neon light or by sunlight. It would not make sense to call the percept of “blue” an illusion_d under neon light when the ambient lighting is such that the object mainly reflects wave lengths of say 500 nm and to call it veridical when it is lit by sunlight such that the domin-

ant wavelength is 450 nm. In both cases, the object appears blue. We cannot determine in principle which of the two cases deserves the name *illusion_d*, if any, or if both deserve to be called *illusion_d*. In contrast, when the two cases are juxtaposed, an *illusion_m* becomes manifest. In Figure 2, the center inner square surrounded by red on the left and the outer squares surrounded by yellow on the right are of an identical color, as becomes manifest when occluding the surrounds. Thus, *illusion_m* becomes apparent, but *illusion_d* cannot be defined in any meaningful way.

4 Conclusion

In sum, the role of illusions in vision research has historically been very important. The beginnings of experimental psychology have attempted to measure *illusions_d* in terms of the discrepancy or error between physical stimulus and percept. I have attempted to show that this error is neither substantial enough to serve as a definition of illusion, nor particularly fascinating. Instead, *illusions_d* are as ubiquitous as they are typically unnoticed or indeterminate. In contrast, the cases that engage our imagination usually are manifest *illusions_m*. The latter can be defined even in cases where it is not meaningful to speak of *illusion_d*.

References

- Hassenstein, B. & Reichardt, W. E. (1956). Systemtheoretische Analyse der Zeit-, Reihenfolgen- und Vorzeichenauswertung bei der Bewegungsperzeption des Rüsselkäfers *Chlorophanus*. *Zeitschrift für Naturforschung*, 11b, 513-524.
- Hecht, H. (2006). Zeitwahrnehmung als Bewegungswahrnehmung. In N. Mewis & S. Schlag (Eds.) *Zeit* (pp. 61-78). Mainz, GER: Leo-Druck.
- Hecht, H. & Proffitt, D. R. (1991). Apparent extended body motions in depth. *Journal of Experimental Psychology: Human Perception and Performance*, 17 (4), 1090-1103. [10.1037/0096-1523.17.4.1090](https://doi.org/10.1037/0096-1523.17.4.1090)
- (1995). The price of expertise: Effects of experience on the water-level task. *Psychological Science*, 6 (2), 90-95.
- McBeath, M. K. & Shepard, R. N. (1989). Apparent motion between shapes differing in location and orientation: A window technique for estimating path curvature. *Perception & Psychophysics*, 46 (4), 333-337.
- Shepard, R. N. & Zare, S. L. (1983). Path-guided apparent motion. *Science*, 220 (4597), 632-634.
- Wertheimer, M. (1912). Experimentelle Studien über das Sehen von Bewegung. *Zeitschrift für Psychologie*, 61 (1), 161-265.

The Neural Organ Explains the Mind

Jakob Hohwy

The free energy principle says that organisms act to maintain themselves in their expected states and that they achieve this by minimizing their free energy. This corresponds to the brain's job of minimizing prediction error, selective sampling of sensory data, optimizing expected precisions, and minimizing complexity of internal models. These in turn map on to perception, action, attention, and model selection, respectively. This means that the free energy principle is extremely ambitious: it aims to explain *everything* about the mind. The principle is bound to be controversial, and hostage to empirical fortune. It may also be thought preposterous: the theory may seem either too ambitious or too trivial to be taken seriously. This chapter introduces the ideas behind the free energy principle and then proceeds to discuss the charge of preposterousness from the perspective of philosophy of science. It is shown that whereas it is ambitious, controversial and needs further evidence in its favour, it is not preposterous. The argument proceeds by appeal to: (i) the notion of inference to the best explanation, (ii) a comparison with the theory of evolution, (iii) the notion of explaining-away, and (iv) a "bio-functionalism" account of Bayesian processing. The heuristic starting point is the simple idea that the brain is just one among our bodily organs, each of which has an overall function. The outcome is not just a defence of the free energy principle against various challenges but also a deeper anchoring of this theory in philosophy of science, yielding an appreciation of the kind of explanation of the mind it offers.

Keywords

Explaining-away | Free energy principle | Functionalism | Inference to the best explanation | Prediction error minimization | Scientific explanation

1 The brain and other organs

Many organs in the body have a fairly specific main function, such as cleaning or pumping blood, producing bile, or digesting. Nothing is ever simple, of course, and all the organs of the body have highly complex, interconnected functional roles. The digestive system involves many different steps; the kidneys help regulate blood pressure; while the heart changes the way it pumps in a very complex and context-dependent manner. Experts in different areas of human biology have a wealth of knowledge about the morphology and physiology of organs, at multiple levels of description. For example, much is known about what cellular and molecular processes occur as the kidneys filter blood, or as food is digested. Knowledge about the functions of organs is not yet complete, but there is reasonable agreement about

the overall picture—namely, which organs have what function.

But the brain seems different. There is much less agreement about what is its main function, and much less knowledge about how it fulfills the various functions attributed to it. Of course, everyone agrees that the brain subserves perception, decision-making, and action—and perhaps that it is the seat of consciousness, self and soul. There is a reasonable degree of knowledge about some aspects of the brain, such as the mechanism behind action potentials, and about what happens when neurons fire. But most would agree that it would be controversial or even preposterous to claim that there is one main function of the brain, on a par with the heart's pumping of blood.

Author

Jakob Hohwy

jakob.hohwy@monash.edu

Monash University
Melbourne, Australia

Commentator

Dominic Harkness

dharkness@uni-osnabrueck.de

Universität Osnabrück
Osnabrück, Germany

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

Yet there is an emerging view that claims that the brain has one overarching function. There is one thing the brain does, which translates convincingly to the numerous other functions the brain is engaged in. This chapter will introduce this idea and will show that, whereas it may be controversial, the idea is not preposterous. It will help us understand better all the things that the brain does, how it makes us who we are, and what we are.

The main version of the idea is labeled the free energy principle, and was proposed by [Karl Friston \(2010\)](#). It unifies and develops a number of different strands of thinking about the brain, about learning, perception and decision-making, and about basic biology. The principle says that biological organisms on average and over time act to minimize free energy. Free energy is the sum of prediction error, which bounds the surprise of the sensory input to the system. Put one way, it is the idea that brains are hypothesis-testing neural mechanisms, which sample the sensory input from the world to keep themselves within expected states. Generalizing greatly, one might say that, just as the heart pumps blood, the brain minimizes free energy.

Before moving on to introduce and defend this idea, it will be useful to explain why the analogy to the functions of other organs is apt. Once a function is identified it serves as a unifying, organizing principle for understanding what the organ does. For example, even though the heart acts very differently during rest and exercise, it still pumps blood. Similarly, even though the brain acts very differently during the awake state and during sleep it still minimizes free energy. Taking such a general approach therefore helps to provide a unified account of the brain.

Related to this, there is a type of objection that will have little bite on the organ-focused account of the brain. To see this, consider again the heart. The heart pumps blood, and this function is realized in part by the way the contraction of the heart muscle occurs—a process that depends on intricate ion flows across heart cell membranes. One should not object to the notion that the heart pumps blood by referring to the fact that what happens in the heart is an intricate cellular ion flow. This is so even

though one might be able to understand much about the heart just by being told the cellular and molecular story. The story about the function and the story about a level of realization of that function are not in conflict with each other. Similarly, one cannot object to the free energy principle by pointing to facts about what the brain does (e.g., what happens as action potentials are generated, or as long-term potentiation is instantiated). The reason for this is that those low-level processes might be ways of realizing free energy minimization. At best such objections are calls for explanatory work of the sort “how can the generation of action potentials be realizations of free energy minimization?”

These two points together suggest that the functional, organ-based account of the brain is reductionist in two ways (familiar from discussions in philosophy of science). On the one hand it seeks to reduce all the different things the brain does to one principle, namely free energy minimization. This is a kind of theory reduction, or explanatory unification. It says that one theory explains many different things. On the other hand, it is consistent with a kind of metaphysical reduction where the overall function is in the end realized by a set of basic physical processes. Here, mental function is fully physical and fully explained by free energy minimization. It is interesting to note that no one would object to such a two-fold reductionism for the heart and other organs, yet it is controversial or even preposterous to do so for the organ that is the brain. For these reasons, it is useful to keep in mind the simple idea that the brain is also an organ. Much of the discussion in this chapter revolves around these two reductive aspects: how can the free energy principle *explain everything*? And can it provide the *functional* scaffolding that would allow realization by brain activity?

2 Minimizing free energy (or average prediction error minimization)

Consider the following very broad, very simple, but ultimately also very far-reaching claim: the brain’s main job is to maintain the organism within a limited set of possible states. This is a

fairly trivial claim, since it just reflects that there is a high probability of finding a given organism in some and not other states, combined with the obvious point that the organism's brain, when in good working order, helps explain this fact. It is the brain's job to prevent the organism from straying into states where the organism is not expected to be found in the long run. This can be turned around such that, for any given organism, there is a set of states where it is expected to be found, and many states in which it would be surprising to find it. This is surely an entirely uncontroversial observation: we don't find all creatures with equal probability in all possible states (e.g., in and out of water). Indeed, since an organism's phenotype results from the expression of its genes together with the influence of the environment, we might define the phenotype in terms of the states we expect it to be found in, on average and over time: different phenotypes will be defined by different sets of states. This way of putting it then defines the brain's job: it must keep the organism within those expected states. That is, the brain must keep the organism out of states that are surprising given the organism it is—or, in general, the brain must minimize surprise (Friston & Stephan 2007).

Here surprise should not be understood in commonsense terms, in the way that a surprise party, say, is surprising. "Surprise" is technically surprisal or self-information, which is a concept from information theory. It is defined as the negative log probability of a given state, such that the surprise of a state increases the more improbable it is to find the creature in that certain state (in this sense a fish out of water is exposed to a lot of surprise). Surprise is then always relative to a model, or a set of expectations (being out of water is not surprising given a human being's expectations). States in which an organism is found are described in terms of the causal impact from the environment on the organism (for example, the difference to the fish between being in water and being out of water). This, in turn, can be conceptualized as the organism's sensory input, in a very broad sense, including not just visual and auditory input but also important aspects of sensation like ther-

moreception, proprioception, and interoception. Surprising states are then to be understood as surprising sensory input, and the brain's job is to minimize the surprise in its sensory input—to keep the organism within states in which it will receive the kind of sensory input it expects.

To be able to use this basic idea about the brain's overall function in an investigation of all the things minds do we need to ask how the brain accomplishes the minimization of surprise. It cannot assess surprise directly from the sensory input because that would require knowing the relevant probability distribution as such. To do this it would need to, impossibly, average over an infinite number of copies of itself in all sorts of possible states in order to figure how much of a surprise a given sensory input might be. This means that to do its job, the brain needs to do something else; in particular it must harbor and finesse a model of itself in the environment, against which it can assess the surprise of its current sensory input. (The model concerns expected sensory states, it is thus a model of the states of the brain, defined by the sensory boundary in both interoceptive and exteroceptive terms, see Hohwy 2014.)

Assume then that the brain has a model—an informed guess—about what its expected states are, and then uses that model to generate hypotheses that predict what the next sensory input should be (this makes it a generative model). Now the brain has access to two quantities, which it can compare: on the one hand the predicted sensory input, and on the other the actual sensory input. If these match, then the model is a good one (*modulo* statistical optimization). Any difference between them can be conceived as prediction error, because it means that the predictions were erroneous in some way. For example, if a certain frequency in the auditory input is predicted, then any difference from what the actual auditory input turns out to be is that prediction's error.

The occurrence of prediction error means the model is not a good fit to the sensory samples after all, and so, to improve the fit, the overall prediction error should be minimized. In the course of minimizing prediction error, the brain averages out uncertainty about its model,

and hence implicitly approximates the surprise. It is guaranteed to do this by minimizing the divergence between the selected hypothesis and the posterior probability of the hypothesis given the evidence and model. The guarantee stems from the facts that this is a Kullback-Leibler divergence (KL-divergence) which is always zero (when there is no divergence) or positive (when there is prediction error), and which therefore creates an upper bound on the surprise—minimizing this bound will therefore approximate surprise.

The key notion here is that the brain acts to maintain itself within its expected states, which are estimated in prediction error minimization. This is known as the free energy principle, where free energy can be understood as the sum of prediction error (this and the following is based on key papers, such as [Friston & Stephan 2007](#), [Friston 2010](#), as well as introductions in [Clark 2013](#) and [Hohwy 2013](#)). Prediction error minimization itself instantiates probabilistic, Bayesian inference because it entails that the selected hypothesis becomes the true posterior, given evidence and model. On this view, the brain is a model of the world (including itself) and this model can be considered the agent, since it acts to maintain itself in certain states in the world.

3 Varieties of prediction error minimization

The central idea here is that, on average and over the long run, surprising states should be avoided, or, prediction error should be minimized. Prediction error minimization can occur in a number of ways, all familiar from debates on inference to the best explanation and many descriptions of scientific, statistical inference.

First, the model parameters can be revised in the light of prediction error, which will gradually reduce the error and improve the model fit. This is perception, and corresponds to how a scientist seeks to explain away surprising evidence by revising a hypothesis. This perceptual process was alluded to above.

Slightly more formally, this idea can be expressed in terms of the free energy principle in

the following terms. The free energy (or sum of prediction error) equals the negative log probability of the sensory evidence, given the model (the surprise) + a KL-divergence between the selected hypothesis (the hypothesis about the causes of the sensory input, which the system can change to change the free energy), and the true posterior probability of the hypothesis given the input and model. Since the KL-divergence is never negative, this means that the free energy will bound (be larger than) the surprise. Therefore, the system just needs to minimize the divergence to approximate the surprisal.

Second, the model parameters can be kept stable and used to generate predictions—in particular, proprioceptive predictions, which are delivered to the classic reflex arcs and fulfilled there until the expected sensory input is obtained. This is action, and corresponds to how a scientist may retain a hypothesis and control the environment for confounds until the expected evidence obtains. Since action is prediction error minimization with a different direction of fit, it is labeled active inference.

Slightly more formally (and still following Friston), this notion of action arises from another reorganization of the free energy principle. Here, free energy equals complexity minus accuracy. Complexity may be taken as the opposite of simplicity, and is measured as a KL-divergence between the prior probability of the hypothesis (i.e., before the evidence came in) and the hypothesis selected in the light of the evidence. Intuitively, this divergence is large if many changes were made to fit the hypothesis—that is, if the hypothesis has significant complexity compared to the old hypothesis. Accuracy is the surprise about the sensory input given the selected hypothesis—that is, how well each hypothesis fits the input. Free energy is minimized by changing the sensory data, such that accuracy increases. If the selected hypothesis is not changed, then this amounts to sampling the evidence selectively such that it becomes less surprising. This can only happen through action, where the organism re-organizes its sensory organs or whole body, or world, in such a way that it receives the expected sensory data (e.g., holding something closer in order to smell it).

There are further questions one must ask about action: how are goals chosen and how do we work out how to obtain them? The free energy principle can be brought to bear on these questions too. In a very basic way, our goals are determined by our expected interoceptive and proprioceptive states, which form the basis of homeostasis. If we assume that we can approximate these expected states, as described above, what remains is a learning task concerning how we can maintain ourselves in them. This relies on internal models of the world, including, crucially, modeling how we ourselves, through our action, impact on the sensory input that affects our internal states. Further, we need to minimize the divergence between, on the one hand, the states we can reach from a given point and, on the other, the states we expect to be in. Research is in progress to set out the details of this ambitious part of the free energy program.

Third, the model parameters can be simplified (cf. complexity reduction), such that the model is not underfitted or overfitted, both of which will generate prediction error in the long run. This corresponds to Bayesian model selection, where complexity is penalized, and also to how a scientist will prefer simpler models in the long run even though a more complex model may fit the current evidence very well. The rationale for this is quite intuitive: a model that is quite complex is designed to fit a particular situation with particular situation-specific, more or less noisy, interfering factors. This implies that it will generalize poorly to new situations, on the assumption that the world is a fairly noisy place with state-dependent uncertainty. Therefore, to minimize prediction error in the long run it is better to have less complex models. Conversely, when encountering a new situation, one should not make too radical changes to one's prior model. One way to ensure this is to pick the model that makes the least radical changes but still explains the new data within expected levels of noise. This is just what Bayesian model selection amounts to, and this is enshrined in the formulations of the free energy principle. A good example of this is what happens during sleep, when there is no trustworthy sensory input and the brain instead

seems to resort to complexity reduction on synthetic data (Hobson & Friston 2012).

Fourth, the hypotheses can be modulated according to the precision of prediction error, such that prediction error minimization occurs on the basis of trustworthy prediction error; this amounts to gain control, and functionally becomes attention. This corresponds to the necessity for assessment of variance in statistical inference, as well as to how a scientist is guided by, and seeks out, measurements that are expected to be precise more than measurements that are expected to be imprecise.

Precision optimization is attention because it issues in a process of weighting some prediction errors more than others, where the weights need to sum to one in order to be meaningful. Hence, peaks across the prediction error landscape reflect both the magnitude of the prediction error per se and the weight given to that error based on how precise it is expected to be. This moves the prediction error effort around, much like one would expect the searchlight of attention to move around.

Within this framework, there is room for both endogenous and exogenous attention. Endogenous attention is top-down modulation of prediction error gain based on learned patterns of precision. Exogenous attention is an intrinsic gain operation on error units, sparked by the current signal strength in the sensory input; this is based on a very basic learned regularity in nature, namely that strong signals tend to have high signal to noise ratio—that is, high precision.

In all this, there is a very direct link between perception, action, and attention, which will serve to illustrate some of the key characteristics of the framework. In particular, expected precision drives action such that sensory sampling is guided by hypotheses that the system expects will generate precise prediction error. A very simple example of this is hand movement. For hand movement to occur, the system needs to prioritize one of two competing possible hypotheses. The first hypothesis is that the hand is *not* moving, which predicts a particular kind of (unchanging) proprioceptive and kinesthetic input; the second hypothesis is (the

false one) that the hand *is* moving, which predicts a different (changing) flow of proprioceptive and kinesthetic input. Movement will only occur if the second hypothesis is prioritized, which corresponds to the agent harboring the belief that the hand is actually moving. If this belief wins, then proprioceptive predictions are passed to the body, where classic reflex arcs fulfill them. Movement is then conceived as a kind of self-fulfilling prophecy.

A crucial question here is how the actually false hypothesis might be prioritized, given that the actually true hypothesis (that the agent is not moving) has evidence in its favor (since the agent is in fact not moving). Here expected precisions play a role, which means that action essentially turns into an attentional phenomenon: in rather revisionist terms, agency reduces to self-organisation guided by long term prediction error minimization. Hypotheses can be prioritized on the basis of their expected precision: hence if future proprioceptive input is expected to be more precise than current proprioceptive input, the gain on the current input will be turned down, depriving the hypothesis that the agent is not moving of evidence. Now the balance shifts in favor of the actually false hypothesis, which can then begin to pass its predictions to the sensorimotor system. This rather inferential process is then what causes movement to occur. It is an essentially attentional process because acting occurs when attention is withdrawn from the actual input (Brown et al. 2013).

The outstanding issue for this story about what it takes to act in the world is why there is an expectation that future proprioceptive input will be more precise than the current input. One possibility here is that this is based on a prior expectation that exploration (and hence movement) yields greater prediction error minimization gains in the long run than does staying put. Conversely, this is the expectation that the current state will lose its high-precision status over time. Writ large, this is the prior expectation concerning precisions (i.e., a hyperprior), which says that the world is a changing place such that one should not retain the same hypotheses for too long: when the pos-

terior probability of a hypothesis becomes the new prior, it will soon begin to decrease in probability. This is an important point because it shows that the ability to shift attention around in order to cause action is not itself an action performed by a homunculus. Rather, it is just a further element of extracting statistical information (about precisions) from the world.

4 Hierarchical inference and the recapitulating, self-evidencing, slowing brain

A system that obeys the free energy principle minimizes its free energy, or prediction error, on average and over time. It does this through perception, belief updating, action, attention, and model simplification. This gives us the outline of a very powerful explanatory mechanism for the mind. There is reason to think that much of this explanatory promise can be borne out (Clark 2013; Hohwy 2013).

This mechanism shapes and structures our phenomenology—it shapes our lived, experienced world. A good starting point for making good on this idea is the notion of hierarchical inference, which is a cornerstone of prediction error minimization.

Conceive of prediction error minimization as being played out between overlapping pairs of interacting levels of processing in the brain. A pair has a lower level receives input, and a higher level that generates predictions about the input at the lower level. Predictions are sent down (or “backwards”) where they attenuate as well as possible the input. Parts of the input it cannot attenuate are allowed to progress upwards, as prediction error. The prediction error serves as input to a new pair of levels, consisting of the old upper level, which is now functioning as lower input level, and a new upper level. This new pair of levels is then concerned with predicting the input that wasn’t predicted lower down. This layering can then go on, creating in the end a deep hierarchy in our brains (and perhaps a more shallow hierarchy in some other creatures). The messages that are passed around in the hierarchy are the sufficient statistics: predictions and prediction errors concern-

ing (1) the means of probability distributions (or probability density functions) associated with various sensory attributes or causes of sensory input out there in the world, and (2) the precisions (the inverse of variance) of those distributions, which mediate the expected precisions mentioned above.

The hierarchy gives a deep and varied empirical Bayes or prediction error landscape, where prior probabilities are “empirical” in that they are learned and pulled down from higher levels, so they do not have to be extracted *de novo* from the current input. This reliance on higher levels means that processing at one level depends on processing at higher levels. Such priors higher up are called hyperparameters, for expectations of means, and hyperpriors for expectations of precisions.

The key characteristics of the hierarchy are *time and space*. Low levels of the hierarchy deal with expectations at fast timescales and relatively small receptive fields, while higher levels deal with expectations at progressively slower timescales and wider receptive fields. That is, different levels of the hierarchy deal with regularities in nature that unfold over different spatiotemporal scales. This gives a trade-off between detail and time horizon such that low down in the hierarchy, sensory attributes can be predicted in great detail but not very far into the future, and higher in the hierarchy things can be predicted further into the future but in less detail. This is essential to inference because different causal regularities in nature, working at different time scales, influence each other and thereby create non-linearities in the sensory input. Without such interactions, sensory input would be linear and fairly easy to predict both in detail and far into the future. So the temporal organization of the hierarchy reflects the causal order of the environment as well as the way the causes in the world interact with each other to produce the flow of sensory input that brains try to predict.

The structure of the hierarchy in the brain, and thereby the shape of the inferences performed in the course of minimizing prediction error, must therefore mimic the causal order of the world. This is one reason why hier-

archical inference determines the shape and structure of phenomenology, at least to the extent that phenomenology is representational. The way inference is put together in the brain recapitulates the causes we represent in perception. Moreover, this is done in an integrated fashion, where different sensory attributes are bound together under longer-term regularities (for example, the voice and the mouth are bound together under a longer-term expectation about the spatial trajectories of people). This immediately speaks to long-standing debates in cognitive science, concerning for example the binding problem and cognitive penetrability (for which see Chs. 5-6 in Hohwy 2013). Though there is, of course, much more to say about how prediction error minimization relates to phenomenology, so far this suggests that there is some reason to think the austere prediction error minimization machine can bear out its explanatory promise in this regard.

Goals and actions are also embodied in the cortical hierarchy. Goals are expectations of which states to occupy. Actions ensue, as described above, when those expected states, which may be represented at relatively long timescales, can confidently be translated into policies for concrete actions fulfilled by the body. There are some thorny questions about what these goals might be and how they are shaped. One very fundamental story says that our expected states are determined by what it takes to maintain homeostasis. We are creatures who are able to harness vast and deep aspects of the environment in order to avoid surprising departures from homeostasis; though this opportunity comes with the requirement to harbor an internal model of the environment. Reward, here, is then the absence of prediction error, which is controlled by using action to move around in the environment, so as to maintain homeostasis on average and in the long run.

Taking a very general perspective, the brain is then engaged in maintaining homeostasis, and it does so by minimizing its free energy, or prediction error. Minimization of prediction error entails building up and shaping a model of the environment. The idea here is very simple. The better the model is at minimizing

prediction error the more information it must be carrying about the true causes of its sensory input. This means that the brain does its job by recapitulating the causal structure of the world—by explaining away prediction error, the brain is essentially becomes a deeply structured mirror of the world. This representational perspective is entailed by the brain's efforts to maintain itself in a low entropy or free energy state. This means that we should not understand the brain as first and foremost in the business of representing the world, such that it can act upon it—which may be an orthodox way of thinking about what the brain does. Put differently, the brain is not selected for its prowess in representation per se but rather for its ability to minimize free energy. Even though this means representation is not foundational in our explanation of the brain, it doesn't mean that representation is sidelined. This is because we don't understand what free energy minimization is unless we understand that it entails representation of the world. (This formulation raises the issue of the possibility of misrepresentation in prediction error minimization, for discussion see [Hohwy 2013](#), Chs. 7-8.)

The brain can be seen, then, as an organ that minimizes its free energy or prediction error relative to a model of the world and its own expected states. It actively changes itself and actively seeks out expected sensory input in an attempt to minimize prediction error. This means the brain seeks to expose itself to input that it can explain away. If it encounters a change in sensory input that it cannot explain away, then this is evidence that it is straying from its expected states. Of course, the more it strays from its expected states, the more we should expect it to cease to exist. Put differently, the brain should enslave action to seek out evidence it can explain away because the more it does so, the more it will have found evidence for its own existence. The very occurrence of sensory input that its model can explain away becomes an essential part of the evidential basis for the model. This means the brain is self-evidencing ([Hohwy 2014](#)), in that the more input it can explain away, the more it gains evidence for the

correctness of the model and thereby for its own existence.

The notions of recapitulation of the world and of self-evidencing can be captured in an exceedingly simple idea. The brain maintains its own integrity in the onslaught of sensory input by *slowing down* and controlling the causal transition of the input through itself. If it had no means to slow down the input its states would be at the mercy of the world and would disperse quickly. To illustrate, a good dam-builder must slow down the inflow of water by slowing down and controlling it with a good system of dams, channels, and locks. This dam system must in some sense anticipate the flows of water in a way that makes sense in the long run and that manages flows well on average. The system will do this by minimizing “flow errors”, and it and its dynamics will thereby carry information about—recapitulate—the states of water flow in the world on the other side of the dam. In general, it seems any system that is able to slow the flow of causes acting upon it must be minimizing its own free energy and thereby be both recapitulating the causes and self-evidencing ([Friston 2013](#)).

With these extremely challenging and abstract ideas, the brain is cast as an organ that does one thing only: minimize free energy and thereby provide evidence for its own existence. Just as the heart can change its beat in response to internal and external changes, the brain can change its own states to manage self-evidencing according to circumstances: perceive, act, attend, simplify. The weighting between these ways of minimizing prediction error is determined by the context. For example, it may be that learning is required before action is predicted to be efficient, so perception produces a narrow prediction error bound on surprise before action sets in, conditional on expected precisions; or perhaps reliable action is not possible (which may happen at night when sensory input is so uncertain that it cannot be trusted) and therefore the brain simplifies its own model parameters, which may be what happens during sleep ([Hobson & Friston 2012](#)).

This is all extremely reductionist, in the unificatory sense, since it leaves no other job for

the brain to do than minimize free energy—so that everything mental must come down to this principle. It is also reductionist in the metaphysical sense, since it means that other types of descriptions of mental processes must all come down to the way neurons manage to slow sensory input.

The next sections turn to the question of whether this extreme explanatory and reductionist theory is not only controversial and ambitious but also preposterous.

5 A preposterous principle? Comparing the free energy principle with evolution

One way to curtail the free energy principle is to allow that the idea of a hypothesis-testing mechanism in the brain may be useful for some but *not all* purposes. Thus the idea could explain, say, visual illusions, but not action. Indeed, versions of the idea in this curtailed form have surfaced many times in the history of philosophy of mind, vision science, and psychology (see [Hohwy 2013](#), Introduction). One view would be that evolution very likely has recruited something like hypothesis-testing, such that the brain can represent the world, but that this likely co-exists with many other types of mechanism that the brain makes use of, for good evolutionary reasons. From this perspective, the universal ambition of the free energy principle is preposterous because it goes against the evolutionary perspective of a tinkering, cobbled-together mechanism.

It is possible of course that a limited-use, Bayesian neural mechanism has evolved in this way. There is no strong evidence that there is in fact something like a circumscribed, modular mechanism. For example, Bayes optimal integration seems to work across modalities and types of sensory attributes ([Trommershäuser et al. 2011](#)). On the other hand, there is not yet strong empirical evidence for the ubiquitousness of free energy minimization, though there is emerging evidence of its usefulness for explaining a very surprising range of mental phenomena, from visual perception, illusion, movement, decision, and action.

Speaking more conceptually, the free energy principle is not a theory that lends itself

particularly well to piecemeal, curtailed application. Recall that the principle concerns the very shape and structure of the brain, mirroring as it does the causal structure of the world. The very hierarchical morphology of the organ is shaped by free energy minimization. This means that other neural mechanisms, that are not involved in prediction error minimization, would have to have evolved in a way parasitic on the free energy principle rather than alongside it. In this sense, the free energy principle would, at the very least, lay the foundation for everything else. Against this, it could be said that perhaps parts of the brain are not, strictly speaking, part of hierarchical inference. Perhaps subcortical nuclei have evolved independently of free energy. This is therefore an argument for which empirical evidence would be important: are there areas of the brain that are not best described in terms of prediction error message passing?

Continuing the very general approach, the free energy principle has such generality that it tends to monopolize explanation. To demonstrate this, consider the theory of evolution, which is also an extremely ambitious theory in the sense that it aims to explain all parts of biology with just a few very basic tools. It is conceptually possible to curtail this theory: perhaps it explains only 70% of life, leaving some other mechanism to explain the rest, or perhaps it explains only non-human life, leaving some deity to fully explain us. This kind of curtailed view would of course ignore the mountain of evidence there is for evolution in absolutely all parts of life (a point we will revisit in a moment), but it would also miss something about the kind of theory that the theory of evolution is. It seems that, as an explanation, evolution is so powerful that it would be incredible that something else would be equally able to explain life.

Whereas it cannot be stipulated that the theory of evolution is true universally, it can be argued that if it is true, it is true everywhere. To see this, consider that if incontrovertible evidence was found that evolution does not explain, say, the eight eyes of most spiders, then for most people that would cast aspersions on the theory of evolution in all other areas—even

where it is backed up with overwhelmingly strong evidence. This is not simply to say that some recalcitrant evidence lowers the posterior probability of the theory somewhat, but rather that it would begin to completely undermine the theory. It seems the theory of evolution posits such a fundamental mechanism that anything short of universal quantification would invalidate it.

Perhaps we can describe what goes on in terms of “explaining away” (Pearl 1988). Imagine, for example, that one night the electricity in your house cuts out. You consider two hypotheses: that a possum has torn down the power line to your house, or that the whole neighbourhood has blacked out due to the recent heat wave. Out in the street you see other people checking their fuse boxes and this evidence favours the second hypothesis. Importantly, this evidence considerably lowers the probability of the possum hypothesis even though the two hypotheses could be true together. There is debate about what explaining away really is, but agreement that it exists. Part of what grounds this notion is that our background knowledge of the frequency of events tells us that it would be rather an unusual coincidence if, just as the overall power goes out due to the heat wave, a possum caused the line to go down (unless possums are known to take to power lines during heat waves). In the case of the deity hypothesis and the evolutionary hypothesis, it seems that explaining away is particularly strong. It would be an utterly astounding coincidence if something as fundamental as speciation and adaptation had two coinciding explanations.

After this excursion into philosophy of science, we can return to the free energy principle. Though it still has nothing like the amount of evidence in its favour that evolution has, it seems that if it is true then it too must apply everywhere, and if not then it must be false. There is no middle way. This again seems to relate to explaining away. It would be too much of a coincidence if two explanations both accounted for something as fundamental as the organism’s ability to sustain itself in its expected states. If the principle was directed at only fairly superficial aspects of mentality, such as

the nature of visual illusions, then it would not strongly explain away other theories. But this misrepresents how deep the explanatory target actually is.

The issue was whether the explanatory ambition of the free energy principle can be curtailed, in order to make it seem less preposterous. If it is assumed that explaining away is particularly strong for fundamental rather than superficial explanations, then it appears that a principle as fundamental as the free energy principle cannot be curtailed. If it is believed, then it is believed with maximal scope. It is therefore misguided to think that one can take a divide and conquer approach to the free energy principle.

Of course, this can be taken to cement its preposterousness. If it is a hypothesis designed to be universal, then how can it be anything but preposterous? The immediate answer to this lies in comparing it again to the theory of evolution. This venerable theory must be preposterous in just the same way, but of course it isn’t—it is true. This means that the issue whether the free energy principle is preposterous cannot be decided just by pointing to its explanatory ambition, since this would also invalidate the theory of evolution. Not surprisingly, it must be resolved by considering the evidence in favour of the free energy principle. As mentioned, this does not yet compare to that of the theory of evolution, though it is noteworthy that evidence is coming in, and that it is coming in from research on a comfortably large suite of mental phenomena.

Consider next the question of what happens with existing, competing theories once something like the free energy principle or the theory of evolution begins to gain explanatory force. Existing theories may have considerable evidence in their favour (this may be a theory about a cognitive or perceptual domain, such as attention or illusion); and they may explain away the existing evidence relatively well and therefore have that evidence in their favour (this contrasts with the comparison with the deity hypothesis, which strictly speaking has no evidence in its favour). Nevertheless, once additional, relevant evidence becomes available, ex-

isting theories may begin to lose ground to a new theory, like the free energy principle, even if it as yet has less evidence in its favour. For example, once it is noted that the brain is characterized by plentiful backwards connections, it becomes clear that these must be relevant to phenomena like attention and illusion (for example, disrupting them disrupts attention and illusion). However, if existing theories cannot explain this new evidence, then a new theory can begin to usurp their explanatory job. This means the evidence in their favour begins to wane, even if the new theory is still only enjoying spotty support. Compare again to the electricity blackout example. There might be a very impressive theory of the whereabouts and heat wave-related behavior of possums that very snugly explains the blackout in the house and perhaps other things besides. But the moment we become aware that the whole neighborhood is without electricity, even a poor theory of the blackout that can also address this new evidence (“perhaps it is some central distributor thingamajig that has broken down”) becomes much more attractive than the existing possum theory. New evidence and new theories can very quickly wreak havoc on old, cherished theories. The free energy principle should therefore be expected to usurp the explanatory jobs of existing theories, and thereby challenge them, even if it is still a fairly fledgling theory. Of course, explanatory usurpation depends on acknowledging the occurrence of new evidence, such as the presence of backwards connections in the brain. Perhaps it is no surprise that the free energy principle is beginning to gain ground just as imaging brain science is maturing beyond the phase in which it was concerned mainly with collecting new evidence, and on to a new phase in which researchers consider the theoretical significance of the evidence in terms of both functional specificity and effective connectivity.

6 Predictions, distinctness, fecundity

It will be useful to discuss a concrete example of explanatory contest for the free energy principle. A good example comes from [Ned Block & Susanna Siegel \(2013\)](#) who argue against [Andy](#)

[Clark’s \(2013\)](#) version of the predictive processing framework in a way that pertains to the preceding remarks about explanatory prowess and ambition. In a comparison with an existing theory of attentional effects (proposed by [Marisa Carrasco](#)), they argue first that the predictive framework makes false predictions, and second that it offers no distinctive explanations.

As to the first point, Block and Siegel consider the effect where covert attention to a weak contrast grating enhances its perceived contrast. They argue that this increased contrast should be unexpected and therefore should elicit a prediction error that in turn should be extinguished, thereby annihilating the perceptual effect that the account was meant to explain in the first place. However, their argument does not rely on the correct version of the free energy account of attention. Block and Siegel overlook the fact that attention is itself predictive, in virtue of the prediction of precision. This means that attention enhances the prediction error from the weak grating, which in turn is explained away under the hypothesis that a strong contrast grating was present in that location of visual space. This conception of attention thus does yield a satisfactory account of the phenomenon that they claim cannot be explained (attentional enhancing), and it does not generate the false predictions they suggest ([Hohwy 2013](#)).

Block and Siegel’s second point is more difficult to get straight. They argue that the predictive account offers no explanation of attentional findings, in particular relating to receptive field distortions; they then suggest that the account could adopt the existing theory, which asserts that “representation nodes” have shrinking receptive fields. They continue to argue that since the purported prediction error gain relates to error units in the brain rather than representation nodes, the prediction error account cannot itself generate this explanation. The argument is then that if the prediction processing account simply *borrow*s that explanation (namely the existing explanation in terms of representation nodes), it hasn’t offered anything distinctive. Again, this rests on an incorrect reading of the free energy account: error units

are not insulated from representation units. Error units receive the bottom-up signal and this leads to revision of the predictions generated from the representation units. The outstanding question is how the distortions of receptive fields can be explained within the prediction error account.

This question has been addressed within the predictive coding literature. Thus [Spratling \(2008\)](#), who is a proponent of predictive coding accounts of attention, says (referring to the literature on changing receptive fields to which Block and Siegel themselves appeal) “the [predictive processing] model proposes, as have others before, that the apparent receptive field distortion arises from a change in the pattern of feedforward stimulation received by the cell”. That is, increased gain explains the distortion of the receptive field.

In fact, one might speculate that the predictive processing story makes perfect sense of the existence of modulable receptive fields. The receptive field of a given representational unit would, that is, be a function of the prediction error received from below, where—as described earlier—lower levels operate at smaller spatiotemporal scales. To give a toy illustration, assume that a broad receptive field would receive an equal amount of error signal from ten lower units each with smaller receptive fields, whereas a narrow receptive field receives error only from two such units. For the broad receptive field, if the gain on error from lower unit numbers one and two increases due to attention, then the gain on the other eight units decreases (since weights sum to one). Now, the hitherto broad receptive field mainly receives error from two lower units, so its receptive field has automatically shrunk. Attentional effects thus track the effects of expected precisions.

Here a more specific point can be made about Block and Siegel’s argument. The predictive processing account of attention can potentially offer a distinctive explanation of rather finegrained attentional findings. There is also reason to think that this explanation has more promise than existing theories. This is because the existing theories help themselves to the notion of ‘representational nodes’ whereas the free

energy principle explains what these are, what they do, and how they connect with other nodes. Moreover, the prediction error account can deal very elegantly for key receptive field properties ([Rao & Ballard 1999](#); [Harrison et al. 2007](#)).

This seems to be a good example of the situation outlined earlier with respect to the contest between the free energy principle and existing theories. The free energy principle can explain more types of evidence, under a more unificatory framework, and this immediately begins to undermine existing theories. Specifically, the theory that has no role for prediction error in receptive field modulation and activation only in representation nodes is explained away, even if it has significant evidence in its favour.

Underlying this story, there are some larger issues in the philosophy of science. One issue concerns the role of unification in explanation ([Kitcher 1989](#)). This is the idea that there are explanatory dividends in explanations that unify a variety of different phenomena under one theory. Obviously the free energy principle is a strong, ambitious unifier (perception, action, and attention all fall under the principle). Whereas there is discussion about whether this in itself adds to its explanatory ability as such, the ability to unify with other areas of evidence is part of what makes an explanation *better* than others. Noting this aspect of the free energy principle therefore supports it, in an inference to the best explanation ([Lipton 2004, 2007](#)). Confronted with a piecemeal explanation of a phenomenon and a unificatory explanation of the same phenomenon, the inference to the latter is stronger. There may be some difficult assessments concerning which explanation best deals with the available evidence. In the case discussed above, the free energy principle can explain less of the attention-specific evidence than the piecemeal explanation, but on the other hand it can explain more kinds of evidence, it provides explanatory tools that are better motivated (roles of representation and error 725 units), and it offers a more unifying account overall.

A second issue from the philosophy of science, in particular concerning inference to the

best explanation, is the fecundity of an explanation, which is regarded as a best-maker. The better an explanation is at generating new predictions and ways of asking research questions, the stronger is the inference in its favour. Whereas this is not on its own a decider, it is an important contributor to the comparison of explanatory frameworks. Block and Siegel also seem to suggest that the predictive framework has nothing new to offer, or at least very little compared to existing (piecemeal) theories. Their example of a piecemeal theory is Carrasco's impressive work on attention, which has proven extraordinarily fecund, leading to a series of discoveries about attention. Assessing which theory is the more fecund is difficult, however, and involves considerations of unification. The free energy principle, as described above, does not posit any fundamental difference between perception and action. Both fall out of different re-organisations of the principle and come about mainly as different directions of fit for prediction error minimization (Hohwy 2013, 2014). This means that optimization of expected precisions, and thereby attention, must be central to action as well as to perception. This provides a whole new (and thus fecund) source of research questions for the area of action, brought about by viewing it as an attentional phenomenon. Important modeling work has been done in this regard (Feldman & Friston 2010), age-old questions (such as our inability to tickle ourselves) have been re-assessed (Brown et al. 2013), and new evidence concerning self-tickle has been amassed (Van Doorn et al. 2014). Theoretically, this has led to the intriguing idea that action occurs when attention is withdrawn from current proprioceptive input (described above). This idea points to a fully integrated view of attention, where attention is ubiquitous in brain function (with deep connections to consciousness, Hohwy 2012).

There is thus fecundity on both sides of this debate. It is difficult to conclusively adjudicate which side is more fecund, in part because the new research questions are in different areas and with different theoretical impact. It is surprising to be told that too much attention can undermine acuity—which is an example

from Block and Siegel—but it is also surprising to be told that action is an attentional phenomenon.

The third issue from the philosophy of science concerns theory subsumption. It would be very odd if the explanations associated with the free energy principle (e.g., that attention is optimization of expected precision) completely contradicted all existing, more piecemeal explanations of attention. It should be expected that explanations of attention have some overlap with each other, as they are explaining away overlapping bodies of evidence. Indeed, the free energy explanation seems to subsume elements of biased competition theories of attention, as well as elements of Carrasco's theory, as seen above. This raises the question of to what extent a new theory, like the free energy principle's account of attention, really contributes a new and better understanding, especially if it carries within it elements of older theories. One way to go about this question again appeals to inference to the best explanation. The new and the old theories overlap in some respects, but they differ in respect of further elements of unification, theoretical motivation, broadness, fecundity, and so on. It can be difficult to come up with a scheme for precise assessment of these features, but it seems not unreasonable to say that the free energy principle performs best on at least those further elements of what makes explanations best.

At this stage it is tempting to apply the free energy principle to itself. This is an apt move since the idea of the hypothesis-testing brain arose in comparison with scientific practice (Helmholtz 1867; Gregory 1980). On this view, the point of a good scientific theory is to minimize prediction error as well as possible, on average and in the long run. This imputes an overall weighting of all the very same elements to science as we have ascribed to the brain above: revise theories in the light of evidence, control for confounds by making experimental manipulations, be guided by where highly precise evidence is expected to be found, adopt simple theories that diverge minimally from old theories, and let theories have a hierarchical structure such that they can persist in the face

of non-linearities (due to causal interactions) in the evidence. All of these considerations speak in favour of the free energy principle over piecemeal, existing theories. By absorbing and revising older theories under the hierarchically imposed scientific “hyperparameter” of the free energy principle, it seems a very reasonable weighting of all these aspects can be achieved. For example, aspects of Carrasco’s theory are subsumed, but under revised accounts of its notions of the functional role of representation nodes; due to the hierarchical aspect it is able to account for evidence arising under attentional approaches to action; in addition, this subsumption may be fecund, since we could expect it to lead to new findings in action (for example, a prediction that there will be attentional enhancement in the sensorimotor domain, leading to “illusory action”).

7 The triviality worry

There is a different worry about preposterousness, also related to the issue of evidence. This worry is that the free energy principle is so general that anything the brain does can be construed as minimizing its prediction error. This is most clearly seen once the idea is cast in Bayesian terms. The brain harbours priors about the causes in the environment, and it calculates likelihoods that it combines with the priors to arrive at posterior probabilities for the hypotheses in question. One way to make this story apply to a particular case is to ascertain what is believed and then in a retrodictive fashion, posit priors and likelihoods accordingly unto the brain in question. If this can always be done, then the theory is trivialised by “just-so” stories and explains nothing. It is then preposterous because it pretends to be fundamental but is just trivial.

This triviality worry alerts the defender of the free energy principle to some pitfalls, but it is not a critical worry. To see this, an appeal can again be made to the theory of evolution. It is clear that when described in very general terms, anything can be described as enhancing fitness. For example, in an infamous hoax, Ramachandran gave a ridiculous, just-so adapt-

ationist account of why gentlemen prefer blondes (Ramachandran & Blakeslee 1998), which some reportedly took seriously. Yet, no one serious thinks this invalidates the theory of evolution. The reason is, to repeat, that there is abundant solid, non-trivial evidence in favour of evolution. In other words, the presence of just-so triviality at some level of description can co-exist with non-trivial explanations at the level of detailed, quantifiable evidence. Therefore the free energy principle cannot be invalidated just because it invites just-so stories. Of course, it is then hostage to translation into more precise, constricted applications to various domains, where predictions can be quantified and just-so stories avoided. Though there is nowhere near the same evidence that we have for the theory of evolution, evidence of this sort is becoming available (some is reviewed in Hohwy 2013).

Whereas the triviality worry does not invalidate the free energy principle, it does alert to some pitfalls. In particular, when forming hypotheses and when explaining phenomena in Bayesian terms, priors should not be stipulated independently of other evidence. If there is independent reason for asserting a prior with an explanatory role, then it is less likely that this prior is part of a just-so story. Similarly, discovery and manipulation of priors has a particularly important role in the defence of the free energy principle as applied to perception. For example, there is independent evidence that we expect light to come more or less from above (Adams et al. 2004), that objects move fairly slowly (Sotiropoulos et al. 2011), and that we expect others to look at us (Mareschal et al. 2013). Once established on independent grounds, researchers are better able to appeal to such priors in other explanations. This then helps avoid the just-so pitfall.

The triviality worry was that *everything* we do can be made to fit with the free energy principle. A different worry is that almost *nothing* we do fits with the free energy principle. If the free energy principle basically says the brain is an organ that tries to slow down the causal impact upon it from the world, then why don’t organisms with brains just seek out sensory deprivation such as dark, silent rooms (Friston

et al. 2012)? This dark room problem is aired very often and is natural on first thought when considering prediction error minimization. However, it also rests on a fundamental misreading of the free energy principle. The principle is essentially about maintaining the organism in its expected states, homeostatically defined, on average and in the long run. Locking oneself up in a dark silent room will only produce transitory free-energy minimization, as the demands of the world and the body will not be avoided for long. Soon, action is required to seek food, and soon the local council will come round to switch off the gas. It is much better for the brain to harness the deep model of the world in order to control its movement through the environment and thereby maintain itself more efficiently in its expected states.

Notice that this point harks back to a very basic hyperprior mentioned above—namely that the world is a changing place so that occupying the same state for too long will incur increasing free energy costs. This means that even if you currently have the prior that sensory deprivation is the right strategy for minimizing free energy, and even if this strategy works initially (as it does after a long and stressful day), that prior will decrease in strength as time goes by—leading to action and thus escape from sensory deprivation.

This response to the dark room problem in fact has a parallel in evolutionary theory. It has been argued that the free energy principle is false, essentially because not every action contributes directly to instantaneous prediction-error minimization and, analogously, it could be objected that evolutionary theory is false because not every trait directly contributes to instantaneous fitness. But of course this is a poor objection because fitness is measured over longer timescales and some traits, such as spandrels, contribute indirectly to fitness.

This and the preceding two sections have considered whether the explanatory ambition of the free energy principle is preposterous. By comparing the principle with the theory of evolution, and casting the worry in terms of philosophy of science, it can be seen that the explanatory ambition is not preposterous in and of it-

self. The verdict on the principle must come down to the quality of the explanations it offers and the amount of evidence in its favour. The principle is bound to be controversial, however, because it strongly explains away competing theories.

Of course, there are further issues to explore regarding the analogy between the free energy principle and the theory of evolution, and no doubt the analogy will have its limits. One interesting issue concerns the possibility of theory revision and thereby the possibility that the original statement of a theory is strictly speaking, false, even if it is one of those theories with extreme explanatory scope. The notion of natural selection as the only mechanism behind evolution is, for example, put under pressure by the discovery of genetic drift. This has led to revision of the theory of evolution, to encompass drift. Could something similar happen to the free energy principle, or is it in effect so ambitious that it is unrevisable? Conversely, is there any conceivable evidence that could falsify the current version of the theory in a wholesale fashion, rather than the piecemeal, detailed fashion discussed above? There are various answers available here, all of which reflect the peculiar theory emerging from the free energy principle.

First, the current form of the principle itself results from a long series of revisions of the basic idea that the brain engages in some kind of inference. Helmholtz' and Ibn Al Haytham's original ideas (reviewed briefly in Hohwy 2013) have been greatly revised, particularly in response to the mathematical realisation that the inversion of generative models presents an intractable problem, thus calling for variational Bayesian approaches to approximate inference. These developments occurred partly in concert with the empirical discovery that the brain, as mentioned above, is characterized by massive backwards connectivity. It is then not unreasonable to say that older feed-forward versions of computational, information theoretical (e.g., infomax) theories of cognition constitute earlier versions of the free energy principle and that the latter is a revision in the light of formal and empirical discovery. The analogy with theory of

evolution can thus be maintained in at least this backwards-looking respect.

Second, a more forward-looking example concerns the nature of the backwards connectivity in the brain. The free energy principle deems these descending signals *predictions*, but crucially it needs them to be of two kinds, namely predictions of the means of the underlying level's representations, and, as mentioned briefly above, predictions of the precisions of the underlying representations (thus encompassing sufficient statistics). There is some direct and some circumstantial evidence in favour of this dual role for descending signals, but the empirical jury is still out. Should it be found that descending signals do not mediate expected precisions, this would falsify the free energy principle. Notice that this falsification would be specific to the free energy principle, since the element of expected precisions is not found in some of the much broader theories in the academic marketplace that seem to countenance a predictive element in cognition. Notice also that a failure to identify top-down expectations of precision would amount to a wholesale falsification of the principle, since these “second-order” expectations are crucial not only for perception but also for action and action initiation (as explained above).

Third, and speaking much more generally, the principle would be falsified if a creature was found that did not act at all to maintain itself in a limited set of states (in our changing world). Such a creature should not on average and over time change its model parameters or active states and yet it would be able to prevent itself from being dispersed with equal probability among all possible states. This is a clear notion of a strong falsifier, and it speaks to the beauty of the free energy principle since it showcases its deep link between life and mind. However, it is not a very feasible falsifier because there is significant doubt that we would classify such a ‘creature’ as being alive or being a creature at all. Consider, for example, that a simple rock would serve as a falsifier in this sense since it is maintained on average and over the long run (that is, its states do not immediately disperse). One possibility here ([Friston](#)

[2013](#)) is to require that the scope is restricted to creatures that are space-filling, that is, who visits the individual states making up their overall set of expected states. A falsifier would then be a creature that manages to be space-filling but who does not manage this by changing its internal and active states via variational Bayes.

One nice question, in all of this, is whether the theory of evolution and the free energy principle can co-exist—and if so, how. This is a substantial issue, and a pertinent one, since both theories are fundamental and pertain to some of the same aspects—such as morphology, phenotypes, and life. Here is not the place to try to answer this interesting question, though inevitably some initial moves are made that might start to integrate them.

8 How literally is the brain Bayesian?

Bayes' rule is difficult to learn and takes considerable conscious effort to master. Moreover, we seem to flout it with disturbing regularity ([Kahneman et al. 1982](#)). So it is somewhat hard to believe that the brain unconsciously follows Bayes' rule. This raises questions about how literally we should think of the brain as a Bayesian hypothesis-tester. In blog correspondence, Lisa Bortolotti put the question succinctly:

Acknowledging that prior beliefs have a role in perceptual inference, do we need to endorse the view that the way in which they constrain inference is dictated by Bayes' rule? Isn't it serendipitous that something we came up with to account for the rationality of updating beliefs is actually the way in which our brain unconsciously works?

Part of the beauty of the free energy principle is that even though it begins with the simple idea of an organism that acts to stay within expected states, its mathematical formulation forces Bayesian inference into the picture. Expected states are those with low surprisal or self-information. That is they have high probability given the model (low negative log probability).

These states cannot be estimated directly because that would require already knowing the distribution of states one can be in. Instead it is estimated indirectly, which is where the free energy comes in. Free energy, as mentioned above, is equal to the surprisal plus the divergence between the probability of the hypothesis currently entertained by the brain's states and the true posterior of the hypothesis given the model and the state. This much follows from Bayes' rule itself. This means that if the brain is able to minimize the divergence, then the chosen hypothesis becomes the posterior. This is the crucial step, because a process that takes in evidence, given a prior, and ends up with the posterior probability, as dictated by Bayes, must at least implicitly be performing inference (Friston 2010).

Hence, if the free energy principle is correct, then the brain must be Bayesian. How should this be understood? Consider what happens as the divergence is minimized. Formally this is a Kullback-Leibler divergence (or cross entropy), which measures the dissimilarity between two probability distributions. The KL-divergence can be minimized with various minimization schemes, such as variational Bayes. This plays an important role in machine learning and is used in simulations of cognitive phenomena using the free energy principle. Given the detail and breadth of such simulations, it is not unreasonable to say that brain activity and behavior are describable using such formal methods.

The brain itself does not, of course, know the complex differential equations that implement variational Bayes. Instead its own activity is brought to match (and thereby slow down) the occurrence of its sensory input. This is sufficient to bring the two probability distributions closer because it can only do this if it is in fact minimizing prediction error. This gives a mechanistic realization of the hierarchical, variational Bayes. The brain is Bayesian, then, in the sense that its machinery implements Bayes not serendipitously but necessarily, if it is able to maintain itself in its expected states. (There is discussion within the philosophy of neuroscience about what it means for explanations to be

computational. See papers by Piccinini 2006, Kaplan 2011, Piccinini & Scarantino 2011, Chirimuuta 2014.)

The notion of realization (or implementation, or constitution) is itself subject to considerable philosophical debate. A paradigmatic reading describes it in terms of what plays functional roles. Thus a smoke alarm can be described in terms of its functional role (*i.e.*, what it, given its internal states, does, given a certain input). The alarm has certain kinds of mechanisms, which realize this role. This mechanism may comprise radioactive ions that react to smoke and causes the alarm to sound. The analogy between the smoke alarm and the brain seems accurate enough to warrant the paradigmatic functionalist reading of the way neuronal circuitry implements free energy minimization and therefore Bayes. Perhaps it is in some sense a moot point whether the ions in the smoke alarm “detect smoke” or whether they should merely be described in terms of the physical reactions that happen when they come into contact with the smoke particles. Rather than enter this debate it seems better to return to the point made at the start, when the brain was compared to other organs such as the heart. Here the point was that it is wrong to retract the description of the heart as a blood pump when we are told that no part of the cardiac cells are themselves pumps. The brain is literally Bayesian in much the same sense as the heart is literally a pump.

Behind this conceptual point is a deeper point about what kind of theory the free energy principle gives rise to (the following discussion will be based on Hohwy 2014). As described above, the Bayesian brain is entailed by the free energy principle. Denying the Bayesian brain then requires denying the free energy principle and the very idea of the predictive mind. This is, of course, a possible position that one could hold. One way of holding it is to “go down a level” such that instead of unifying everything under the free energy principle, theories just describe the dynamical causal interactions between brain and world. This would correspond to focusing more on systematic elements in the realization than in the function (looking

at causal interaction between the heart and other parts of the body, and the individual dynamics of the cells making up the heart, rather than understanding these in the light of the heart being a pump). Call this the “causal commerce” position on the brain. Given the extensive and crucial nature of causal commerce between the brain and the world, this is in many ways a reasonable strategy. It seems fair to characterize parts of the enactive cognition position on cognitive science as informed primarily by the causal commerce position (for a comprehensive account of this position, see [Thompson 2007](#); for an account that brings the debate closer to the free energy principle, see [Orlandi 2013](#)).

From this perspective, the choice between purely enactive approaches and inferential, Bayesian approaches becomes methodological and explanatory. One key question is what is accomplished by re-describing the causal commerce position from the more unified perspective of the free energy principle. It seems that more principled, integrated accounts of perception, action, and attention then become available. The more unified position also seems to pull away from many of the lessons of the enactive approach to cognition, because the free energy principle operates with a strict inferential veil between mind and world—namely the sensory evidence behind which hidden causes lurk, which must be inferred by the brain. Traditionally, this picture is anathema to the enactive, embodied approaches, as it lends itself to various forms of Cartesian skepticism, which signals an internalist, secluded conception of mind. A major challenge in cognitive science is therefore to square these two approaches: the dynamical nature of causal commerce between world, body, and brain and the inferential free energy principle that allows their unification in one account. On the approach advocated here, modulo enough empirical evidence, denying that the free energy principle describes the brain is on a par with denying that the heart is a pump. This means that it is not really an option to deny that the brain is inferential. This leaves open only the question of *how* it is inferential.

One line of resistance to subsuming everything under the free energy principle has to do with intellectualist connotations of Bayes. Somehow the idea of the Bayesian brain seems to deliver a too regularized, sequential, mathematical desert landscape—it is like a picture of a serene, computational mechanism silently taking in data, passing messages up and down the hierarchy, and spitting out posterior probabilities. This seems to be rather far from the somewhat tangled mess observed when neuroscientists look at how the brain is in fact wired up. In one sense this desert landscape is of course the true picture that comes with the free energy principle, but there need be nothing serene or regularized about the way it is realized. The reason for this goes to the very heart of what the free energy principle is. The principle entails that the brain recapitulates the causal structure of the world. So what we should expect to find in the brain will have to be approximating the far-from-serene and regularized interactions that occur between worldly causes. Just as there are non-linearly interacting causes in the world there will be convolving of causes in the brain; and just as there are localized, relatively insulated causal “eddies” in the world there will be modularized parameter spaces in the brain.

Moreover, there is reason to think the brain utilizes the fact that the same causes are associated with multiple effects on our senses and therefore builds up partial models of the sensorium. This corresponds to cognitive modules and sensory modalities allowing processing in conditionally independent processing streams, which greatly enhances the certainty of probabilistic inference. In this sense the brain is not only like a scientist testing hypotheses, but is also like a courtroom calling different, independent witnesses. The courtroom analogy is worth pursuing in its own right ([Hohwy 2013](#)), but for present purposes it supports the suggestion that when we look at the actual processing of the brain we should expect a fairly messy tangle of processing streams. ([Clark 2013](#) does much to characterize and avoid this desert landscape but seems to do so by softening the grip of the free energy principle.)

9 Functionalism and biology

So far the free energy principle has been given a functionalist reading. It describes a functional role, which the machinery in the brain realizes. One of the defining features of functionalism is that it allows multiple realization. This is the simple idea that the same function can be realized in different ways, at least in principle. For example, a smoke alarm is defined by its functional role but can be realized in different ways. There is on-going debate about whether something with the same causal profile as the human brain could realize a mind. Philosophers have been fond of imaging, for example, a situation in which the population of Earth is each given a mobile phone and a set of instructions about whom to call and when, which mimics the “instructions” followed by an individual neuron (Block 1976). The question then is whether this mobile phone network would be a mind. Though this is not the place to enter fully this debate, it seems hard for the defender of the free energy principle to deny that, if these mobile phone-carrying individuals are really linked up in the hierarchical message-passing manner described by the equations of the free energy principle, if they receive input from hidden causes, and if they have appropriate active members, then they do constitute a mind.

However, a different issue here is to what extent the free energy principle allows for the kind of multiple realization that normally goes with functionalism. The mathematical formulations and key concepts of the free energy principle arose in statistical physics and machine learning, and hierarchical inference has been implemented in computer learning (Hinton 2007). So there is reason to think that prediction error minimization can be realized by computer hardware as well as brainware. There is also reason to think that within the human brain the same overall prediction error minimization function can be realized by different hierarchical models. Slightly different optimizations of expected precisions would determine the top-down vs. bottom-up dynamics differently, but may show a similar ability to minimize prediction error over some timeframes. Different weightings of low

and high levels in the hierarchy can lead to the same ability to minimize prediction error in the short and medium term. This is similar to how a dam can be controlled with many small plugs close to the dam wall, or by fewer connected dam locks operating at longer timescales further back from the dam wall. In some cases, such different realizations may have implications for the organism over the long run, however (for example, building locks in a dam may take time, and thus allow flows in the interim; whereas many small plugs prevent flows in the short run but may be impractical in the long run). Such differences may show up in our individual differences in perceptual and active inference (for an example, see Palmer et al. 2013), and may also be apparent in mental illness (Hohwy 2013, Ch. 7).

Functionalist accounts of the mind are widely discussed in the philosophical literature, and there are various versions of it. A key question for any functionalism is how the functional roles are defined in the first instance (for an overview see Braddon-Mitchell & Jackson 2006). Some theories—psychofunctionalism or empirical functionalism—posit that functional roles should be informed by best empirical science (“pain is caused by nociceptor activation... etc.”). The consequence is that their domain is restricted to those creatures for whom that empirical science holds. Other theories—commonsense functionalism—begin with conceptual analysis and use that to define the functional roles (“pain is the state such that it is caused by bodily damage, gives rise to pain-avoidance behavior, and relates thus and so to internal states...”). The consequence of taking the commonsense approach is that such functionalisms apply widely, including to creatures science has never reached, in so far as they have something realizing that functional role.

There are some nice questions here about what we should really say about creatures with very different realizations of the same functions (e.g., “Martian pain”), and creatures with very similar realizations but different functions (e.g., “mad pain”; see Lewis 1983). Setting those issues aside for the moment, one question is which kind of functionalism goes with the free

energy principle. There is no straightforward answer here, but one possibility is that it is a kind of “biofunctionalism”, where the basic functional role is that of creatures who manage to maintain themselves within a subset of possible states (in a space-filling or active manner) for a length of time. Any such creature must be minimizing its free energy and hence engaging in inference and action. It is biological functionalism because it begins by asking for the biological form—the phenotype—of the candidate creature.

This is an extremely abstract type of functionalism, which allows considerable variation amongst phenotypes and hence minds. For example, it has no problem incorporating both Martians and madmen in so far as they maintain themselves in their expected states. It will however specify the mental states of the organism when it becomes known in which states it maintains itself. This follows from the causal characterization of sensory input, internal states, and active output that fully specify a prediction error minimizing mechanism. Once these states are observed, the states of the system can be known too, and the external causes rendered uninformative (i.e., the sensory and active states form a Markov blanket; [Friston 2013](#)).

What drives biofunctionalism is not species-specific empirical evidence, as in psychofunctionalism. And it does not seem to be commonsense conceptual analysis either. Rather, it begins with a biological, statistical observation that is as basic as one can possibly imagine—namely that creatures manage to maintain themselves within a limited set of states. As seen at the very start of this paper, this defines a probability density for a given creature, which it must approximate to do what it does. For an unsupervised system, it seems this can only happen if the organism minimizes its free energy and thereby infers the hidden causes of its sensory input, and then acts so as to minimize its own errors. This is an empirical starting point at least in so far as one needs to know many empirical facts to specify which states a creature occupies. But it is, arguably, also a conceptual point in so far as one hasn’t under-

stood what a biological creature is if one does not associate it at least implicitly with filling some specified subset of possible states.

The upshot is that the free energy principle sits well with a distinct kind of functionalism, which is here called biofunctionalism. It remains an open question how this would relate to some versions of functionalism and related views, such as teleosemantics ([Neander 2012](#)), which relies on ideas of proper function, and information theoretical views ([Dretske 1983](#)). The biofunctionalism of the free energy principle seems to have something in common with those other kinds of positions though it has no easy room for the notion of proper function and it doesn’t rely on, but rather entails, information theoretical (infomax) accounts.

Setting aside these theoretical issues, note that biofunctionalism has a rather extreme range because it entails that there is Bayesian inference even in very simple biological organisms in so far as they minimize free energy. This includes for example *E. coli* that with its characteristic swimming-tumbling behavior, maintains itself in its expected states. And it includes us, who with our deeper hierarchical models maintain ourselves in our expected states (with more space-filling and for longer than *E. coli*). Of course, one might ask where, within such a wide range of creatures, we encounter systems that we are comfortable describing as minds—that is, as having thought, as engaging in decision-making and imagery, and not least as being conscious. This remains a challenge for the free energy principle, just as it is a challenge for any naturalist theory of the mind to specify where, why, and how these distinctions between creatures arise.

10 The neural organ can explain the mind

The brain is an organ with a function, namely to enable the organism to maintain itself in its expected states. According to the free energy principle, this is to say that it minimizes prediction error on average and over the long run. This is a controversial idea, with extreme explanatory ambition. It might be considered not only controversial but also preposterous. But

the philosophy of science-based discussions above have sought to show that it is not in fact preposterous. The different ways in which it might be preposterous either do not apply, misunderstand the principle, or would also apply to the paradigmatically non-preposterous theory of evolution. The free energy principle yields a theory that should, indeed, strongly explain away competing theories. The free energy principle is an account that displays a number of explanatory virtues such as unification and fecundity. It is therefore not reasonable to detract from the principle by claiming it is preposterous or too ambitious. Scientifically speaking, what remains is to assess the evidence for and against the free energy principle and consider how, more specifically, it explains our mental lives (a task I undertake in Hohwy 2013). Speaking in terms of philosophy of mind, there remain questions about what type of functionalist theory the free energy principle is, how it performs vis-à-vis traditional questions about functionalism and the realizers of functional roles, and, finally, some more metaphysical questions about what it says about the nature of the mind in nature. None of these philosophical issues are apparently more damning for the free energy principle than they are for other, previously proposed accounts of the nature of the mind, and there is reason to think that with the free energy principle a new suite of answers may become available.

References

- Adams, W. J., Graf, E. W. & Ernst, M. O. (2004). Experience can change the ‘light-from-above’ prior. *Nature Neuroscience*, 7 (10), 1057-1058. [10.1038/nn1312](https://doi.org/10.1038/nn1312)
- Block, N. (1976). Troubles with functionalism. *Minnesota Studies in the Philosophy of Science*, 9, 261-325.
- Block, N. & Siegel, S. (2013). Attention and perceptual adaptation. *Behavioral and Brain Sciences*, 36 (3), 205-206. [10.1017/S0140525X12002245](https://doi.org/10.1017/S0140525X12002245)
- Braddon-Mitchell, D. & Jackson, F. (2006). *The philosophy of mind and cognition: An introduction*. London, UK: Wiley-Blackwell.
- Brown, H., Adams, R. A., Parees, I., Edwards, M. & Friston, K. (2013). Active inference, sensory attenuation and illusions. *Cognitive Processing*, 14 (4), 411-427. [10.1007/s10339-013-0571-3](https://doi.org/10.1007/s10339-013-0571-3)
- Chirimuuta, M. (2014). Minimal models and canonical neural computations: The distinctness of computational explanation in neuroscience. *Synthese*, 191 (2), 127-153. [10.1007/s11229-013-0369-y](https://doi.org/10.1007/s11229-013-0369-y)
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral & Brain Sciences*, 36 (3), 181-204. [10.1017/S0140525X12000477](https://doi.org/10.1017/S0140525X12000477)
- Dretske, F. (1983). *Knowledge and the flow of information*. Cambridge, MA: MIT Press.
- Feldman, H. & Friston, K. (2010). Attention, uncertainty and free-energy. *Frontiers in Human Neuroscience*, 4 (215), 1-23. [10.3389/fnhum.2010.00215](https://doi.org/10.3389/fnhum.2010.00215)
- Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11 (2), 127-138. [10.1038/nrn2787](https://doi.org/10.1038/nrn2787)
- (2013). Life as we know it. *Journal of the Royal Society Interface*, 10 (86). [10.1098/rsif.2013.0475](https://doi.org/10.1098/rsif.2013.0475)
- Friston, K., Thornton, C. & Clark, A. (2012). Free-energy minimization and the dark room problem. *Frontiers in Psychology*, 3 (130), 1-7. [10.3389/fpsyg.2012.00130](https://doi.org/10.3389/fpsyg.2012.00130)
- Friston, K. & Stephan, K. (2007). Free energy and the brain. *Synthese*, 159 (3), 417-458. [10.1007/s11229-007-9237-y](https://doi.org/10.1007/s11229-007-9237-y)
- Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society B, Biological Sciences*, 290 (1038), 181-197. [10.1098/rstb.1980.0090](https://doi.org/10.1098/rstb.1980.0090)
- Harrison, L. M., Stephan, K. E., Rees, G. & Friston, K. J. (2007). Extra-classical receptive field effects measured in striate cortex with fMRI. *NeuroImage*, 34 (3), 1199-1208. [10.1016/j.neuroimage.2006.10.017](https://doi.org/10.1016/j.neuroimage.2006.10.017)

- Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences*, 11 (10), 428-434. [10.1016/j.tics.2007.09.004](#)
- Hobson, J. A. & Friston, K. J. (2012). Waking and dreaming consciousness: Neurobiological and functional considerations. *Progress in Neurobiology*, 98 (1), 82-98. [10.1016/j.pneurobio.2012.05.003](#)
- Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology*, 3 (96), 1-14. [10.3389/fpsyg.2012.00096](#)
- (2013). *The predictive mind*. Oxford, UK: Oxford University Press.
- (2014). The self-evidencing brain. *Noûs, Early View*. [10.1111/nous.12062](#)
- Kahneman, D., Slovic, P. & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge University Press.
- Kaplan, D. M. (2011). Explanation and description in computational neuroscience. *Synthese*, 183 (3), 339-373. [10.1007/s11229-011-9970-0](#)
- Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In P. Kitcher & W. Salmon (Eds.) *Scientific explanation* (pp. 410-505). Minneapolis, MN: University of Minnesota Press.
- Lewis, D. (1983). Mad pain and martian pain. In D. Lewis (Ed.) *Philosophical papers, Vol. 1* (pp. 122-130). Oxford, UK: Oxford University Press.
- Lipton, P. (2004). *Inference to the best explanation*. London, UK: Routledge.
- (2007). Précis of Inference to the best explanation. *Philosophy and Phenomenological Research*, 74 (2), 421-423. [10.1111/j.1933-1592.2007.00027.x](#)
- Mareschal, I., Calder, A. J. & Clifford, C. W. G. (2013). Humans have an expectation that gaze is directed toward them. *Current Biology*, 23 (8), 717-721. [10.1016/j.cub.2013.03.030](#)
- Neander, K. (2012). Teleological theories of mental content. *The Stanford encyclopedia of philosophy, Spring 2012 Edition* E. N. Zalta (Ed.) <http://plato.stanford.edu/archives/spr2012/entries/content-teleological/>
- Orlandi, N. (2013). Embedded seeing: Vision in the natural world. *Noûs*, 47 (4), 727-747. [10.1111/j.1468-0068.2011.00845.x](#)
- Palmer, C. J., Paton, B., Hohwy, J. & Enticott, P. G. (2013). Movement under uncertainty: The effects of the rubber-hand illusion vary along the nonclinical autism spectrum. *Neuropsychologia*, 51 (10), 1942-1951. [10.1016/j.neuropsychologia.2013.06.020](#)
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, CA: Morgan Kaufmann Publishers.
- Piccinini, G. (2006). Computational explanation in neuroscience. *Synthese*, 153 (3), 343-353. [10.1007/s11229-006-9096-y](#)
- Piccinini, G. & Scarantino, A. (2011). Information processing, computation, and cognition. *Journal of Biological Physics*, 37 (1), 1-38. [10.1007/s10867-010-9195-3](#)
- Ramachandran, V. S. & Blakeslee, S. (1998). *Phantoms in the brain*. London, UK: Fourth Estate.
- Rao, R. P. N. & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive field effects. *Nature Neuroscience*, 2 (1), 79-87. [10.1038/4580](#)
- Sotiropoulos, G., Seitz, A. R. & Seriés, P. (2011). Changing expectations about speed alters perceived motion direction. *Current Biology*, 21 (21), R883-R884. [10.1016/j.cub.2011.09.013](#)
- Spratling, M. W. (2008). Predictive coding as a model of biased competition in visual attention. *Vision Research*, 48 (12), 1391-1408. [10.1016/j.visres.2008.03.009](#)
- Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind*. Harvard, MA: Harvard University Press.
- Trommershäuser, J., Körding, K. & Landy, M. (Eds.) (2011). *Sensory cue integration*. Oxford, UK: Oxford University Press.
- Van Doorn, G., Hohwy, J. & Symmons, M. (2014). Can you tickle yourself if you swap bodies with someone else? *Consciousness and Cognition*, 23, 1-11. [10.1016/j.concog.2013.10.009](#)
- von Helmholtz, H. (1867). *Handbuch der Physiologischen Optik*. Leipzig, GER: Leopold Voss.

From Explanatory Ambition to Explanatory Power

A Commentary on Jakob Hohwy

[Dominic L. Harkness](#)

The free energy principle is based on Bayesian theory and generally makes use of functional concepts. However, functional concepts explain phenomena in terms of how they should work, not how they in fact do work. As a result one may ask whether the free energy principle, taken as such, can provide genuine explanations of cognitive phenomena. This commentary will argue that (i) the free energy principle offers a stronger unification than Bayesian theory alone (strong unification thesis) and that (ii) the free energy principle can act as a heuristic guide to finding multilevel mechanistic explanations.

Keywords

Active inference | Bayesian enlightenment | Bayesian fundamentalism | Bayesian theory | Free energy | Free energy principle | Functional | Mechanisms | Precision | Prediction errors | Preposterous | Strong unification thesis | Weak unification thesis

Commentator

[Dominic L. Harkness](#)

dharkness@uni-osnabrueck.de

Universität Osnabrück
Osnabrück, Germany

Target Author

[Jakob Hohwy](#)

jakob.hohwy@monash.edu

Monash University
Melbourne, Australia

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

The free energy principle has far-reaching implications for cognitive science. In fact, the free energy principle seeks to explain everything related to the mind. Due to this explanatory ambition, it has been deemed preposterous by researchers. Jakob Hohwy challenges the opponents of the free energy principle and its applications by demonstrating that this framework is everything but preposterous. Rather, he compares the free energy principle with the theory of evolution in biology. The theory of evolution is not discarded due to its unifying power; and

the free energy principle shouldn't be either. In this paper I will present a negative as well as two positive theses: first, the free energy principle will be contrasted to Bayesian theory with regard to the degree of unification they offer. I will argue that the unification resulting from the free energy principle can be regarded as stronger since it attempts to empirically ground its conclusions in the brain via neuroscience and psychology. The negative thesis consists in the suggestion that one major flaw of the free energy principle, taken as such, lies within its ex-

planatory *power*. As a result of being a functional theory, the concepts it employs are also functional. Yet functional concepts, at least when it comes to explaining the brain and cognitive phenomena, do not explain how a certain phenomenon actually works, but rather how it should work. To improve this situation, the second positive thesis of this paper makes use of a suggestion by [Piccinini & Craver \(2011\)](#), namely that functional analyses are mechanism sketches, i.e., incomplete descriptions of mechanisms. In other words, functional concepts (such as precision) must be enriched with mechanistic concepts that include known structural properties (such as “dopamine”) in order to count as a full explanation of a given phenomenon. The upshot of this criticism lies within the free energy principle’s potential to act as a heuristic guide for finding multilevel mechanistic explanations. Furthermore, this paper will not advocate that functional concepts should be fully replaced or eliminated, but that functional and mechanistic descriptions complement each other.

2 The free energy principle

In his article “The Neural Organ Explains the Mind”, [Jakob Hohwy \(this collection\)](#) proposes that the brain, as every other organ in the human body, serves one basic function. Just as one might say that the basic function of the heart is to pump blood through the body or that of the lungs is to provide oxygen, the basic function of the brain is to minimise free energy ([Friston 2010](#)). However, this is a very general claim that does not yet establish how the minimisation of free energy is realised in humans. How is this done?

Very generally, the brain stores statistical regularities from the outer environment or, in other words, it forms an internal model about the causal structure of the world. This model is then used to predict the next sensory input. Consequently, we have two values that can be compared with each other: the predicted sensory feedback and the actual sensory feedback. When perceiving, the brain predicts what its own next state will be. Depending on the accu-

acy of the prediction, a divergence will be present between the predicted and the actual sensory feedback. This divergence is measured in terms of prediction errors. The larger the amount of prediction error, the less accurately the model fits the actual sensory feedback and thus the causal structure of the world. Crucially, the model that fits best, i.e., that which brings forth the smallest amount of prediction error, also determines consciousness. In this framework, free energy amounts to the sum of prediction errors. Thus, minimizing prediction errors always entails the minimisation of free energy.

The minimization of prediction error can generally be achieved in two ways: either the brain can change its models according to the sensory input or, vice versa, it can change the sensory input according to its models. In this scheme the former mode can be seen as veridical perception, whereas the latter can be seen as action, or more formally active inference—the fulfillment of predictions via classic reflex arcs ([Friston et al. 2009](#); [Friston et al. 2011](#)). Furthermore, two other factors play a large role in the minimization of prediction error: first, the precision, or “second-order statistics” ([Hesselmann et al. 2012](#)), which ultimately encodes how “trustworthy” the actual sensory input is. Precision is realised by synaptic gain, and it has been established that the modulation of precision corresponds to attention ([Hohwy 2012](#)). Second, model optimization ensures that models are reduced in complexity in order to account for the largest number of possible states in the long run, i.e., under expected levels of fluctuating noise. For example, sleep has been associated with this type of model optimization ([Hobson & Friston 2012](#)). More detailed descriptions of these four factors, i.e., perception, active inference, precision, and model optimization can be found in Hohwy’s article.

Additionally, models are arranged in a cortical hierarchy ([Mumford 1992](#)). This hierarchy is characterised, as [Hohwy](#) points out ([this collection](#), p. 7), by time and space: models higher up in the hierarchy have a larger temporal scale and involve larger receptive fields than models lower down in the hierarchy, which concern pre-

dictions at fast time scales and involve small receptive fields (p. 7). This hierarchy implies a constant message-passing amongst different levels. Once a sensory signal arrives at the lowest level it is compared to the predictions coming from the next higher level (in this case level two).¹ If prediction errors ensue they are sent to the higher level (still level two). Here they are predicted by the next higher level (now level three). This process goes on until prediction errors are minimised to expected levels of noise.

Now the general scheme of prediction error minimization can be presented: the brain builds models that represent the causal structure of the world. These models are, in turn, used to generate predictions about what the next sensory input might be. The two resulting values, i.e., the predicted and the actual sensory feedback, are continuously compared. The divergence between these two values is the prediction error, or free energy. Since it is the brain's main function to minimise the amount of free energy and therefore prediction error, it will either change its models or engage in active inference. Decisions about which path will be taken depend on the precision of the incoming sensory signal (or prediction error). Signals with high precision are taken to be "trustworthy", and therefore model changes can follow. Low precision signals, however, require further investigation since noise could be the principal factor in an ambiguous input. In addition, models during wakefulness are changed "on-the-fly", thus leading to highly idiosyncratic and complex models. This complexity is reduced, for example during sleep (Hobson & Friston 2012), to increase the generalizability of models, since noise is always present.

3 Bayesian theory and unification

As mentioned above, all this serves the basic function of the brain: the minimization of free energy. This strategy is employed in every aspect of cognition; thus the free energy principle (Friston 2010) is a grand unifying theory. But

from where does the free energy principle derive its unifying power?²

The free energy principle makes use of Bayesian theory, which can be regarded as its foundation. For some years now, Bayesian theory has been applied to many cognitive phenomena, since it may "offer a new interpretation of the goals of cognitive systems, in terms of inductive probabilistic inference [...] can make the assumptions of Bayesian models more transparent than in mechanistically oriented models [...] and may have the potential to explain some of the most complex aspects of human cognition [...]" (Jones & Love 2011, p. 170). Yet Jones & Love (2011) also address the fact that Bayesian theories, although aiming at researching and investigating the human brain and its workings, remain unconstrained by psychology and neuroscience "and are generally not grounded in empirical measurement" (*ibid.*, p. 169). They term this approach "Bayesian Fundamentalism", since it entails that all that is necessary to explain human behaviour is rational analysis. Supporters of this position rely on the mathematical framework of Bayesian theory as the origin of its explanatory power and unification. The positive thesis of Jones & Love (2011) consists in arguing for "Bayesian Enlightenment" that tries to include mechanistic explanation in Bayesian theory. To give more detail, they propose that, rather than following Bayesian Fundamentalism and thus being "logically unable to account for mechanistic constraints on behavior [...] one could treat various elements of Bayesian models as psychological assumptions subject to empirical test" (Jones & Love 2011, p. 184). Similarly, Colombo & Hartmann (2014) argue that although "the Bayesian framework [...] does not necessarily reveal aspects of a mechanism[,] Bayesian unification [...] can place fruitful constraints on causal-mechanical explanation" (Colombo & Hartmann 2014, p. 1).

According to Colombo & Hartmann (2014), many Bayesian theorists falsely equate unification with explanatory power. But Bayesian theories derive their unificatory power

¹ The numerical values for the levels have no scientific relevance. They are used only for illustrative purposes.

² At this point I would like to thank one of the reviewers for her or his substantial advice and constructive comments.

from their mathematical framework. However, just because different cognitive phenomena can be mathematically unified does not entail a causal relationship between them, and nor does the mathematical unification tell us anything about the causal history of these phenomena. However, as will be presented in the next section, explanatory power, at least from a mechanistic point of view, results from investigating structural components and their causal interactions that give rise to a certain phenomenon. For example [Kaplan & Craver \(2011\)](#) write that “[...] the line that demarcates explanations from merely empirically adequate models seems to correspond to whether the model describes the relevant causal structures that produce, underlie, or maintain the explanandum phenomenon” (p. 602). Yet in the case of Bayesian theory—and Bayesian Fundamentalism in particular—, this cannot be achieved, since they “say nothing about the spatio-temporally organized components and causal activities that may produce particular cognitive phenomena [...]” ([Colombo & Hartmann 2014](#), p. 5). But not everything is lost concerning the explanatory role of Bayesian theories. Even if Bayesian theory cannot provide mechanistic explanations, it may nonetheless be beneficial to cognitive science by offering constraints on causal-mechanical explanation ([Colombo & Hartmann 2014](#)).

This brings us to the free energy principle. As noted, the free energy principle is, at its core, a theory that makes use of Bayesian theory; consequently it inherits all of Bayesian theory’s pros and cons. Thus, since unification in the free energy principle is also grounded in its mathematical foundations “[...] the real challenge is to understand how [the free energy principle] manifests in the brain” ([Friston 2010](#), p. 10). With regard to [Jones & Love’s \(2011\)](#) distinction, the free energy principle can be considered to belong to Bayesian Enlightenment, since it attempts to ground its findings in neurobiology and psychology rather than remaining unconstrained by these sciences. Furthermore, due to the fact that the free energy principle integrates neuroscientific findings into its conclusions, it can offer more precise constraints on causal-mechanical explanations than

Bayesian theory alone. For example, the free energy principle tries to incorporate neuroscientific facts about brain structure and its hierarchical organization, or tries to link concepts such as “precision” to neurophysiological phenomena such as “dopaminergic gating” ([Friston et al. 2012](#)).³ The latter example will be presented in greater detail in section 5.

In sum, the free energy principle offers a form of unification that exceeds that offered by Bayesian theory alone. It makes statements about how the free energy principle could be realised in the brain and does not solely rely on its mathematical framework. Thus, one could term the former a “strong unification thesis” (SUT) and the latter a “weak unification thesis” (WUT).

If the free energy principle is true it creates a backdrop against which other theories must be evaluated. This also implies a kind of explanatory monopolization, since “the free energy principle is not a theory that lends itself particularly well to piecemeal” ([Hohwy this collection](#), p. 9). In other words, as Hohwy highlights on many occasions, the free energy principle is an all-or-nothing theory. He compares it to the theory of evolution in biology and states that, just like the free energy principle, “evolution posits such a fundamental mechanism that anything short of universal quantification would invalidate it” (p. 10). Due to this large explanatory ambition, some researchers have described the free energy principle as preposterous. Yet “the issue whether the free energy principle is preposterous cannot be decided just by pointing to its explanatory ambition [...] [but] by considering the evidence in favour of the free energy principle” (p. 11). This is a very important transition, i.e., the switch from explanatory ambition to explanatory power, since, from a mechanistic viewpoint, the former gives no statement about the veridicality of its assumptions, whereas the latter does.

³ However, I’d like to point out that the free energy principle does not make any commitments to one single neuroscientific theory. Rather, it tries to find entities that may realize the free energy principle in the brain; what these entities are remains to be inquired.

In the remainder of this paper, I will argue that one major shortcoming of the free energy principle lies in its explanatory *power*. The main issue to be discussed consists in the fact that most concepts employed in the free energy principle, or in its applications such as predictive coding (Friston 2005; Rao & Ballard 1999) or predictive processing (Clark 2013; Hohwy 2013), are principally functional concepts. Yet, at least in the case of the free energy principle, functional concepts do not hold much explanatory power, since they “describe how things ought to work rather than how they in fact work” (Craver 2013, p. 18). For example, the concept of “precision” represents the amount of uncertainty in the incoming sensory signal that may arise due to noise. Thus the precision of the incoming sensory inputs determines how an agent interacts with its environment next: it can either change its models or its sensory input. Yet, this description holds no commitments as to how precision is realised in the brain; it only describes what effect precision *should* have on a given cognitive system. Therefore the free energy principle seems to be of a normative, rather than descriptive, nature.⁴ On the other hand, there are mechanistic explanations that, according to Craver (2007), can also count as such, since they don’t describe how things should work but how they in fact *do* work.

Yet these two types of epistemic strategies don’t necessarily exclude each other. Here I want to introduce Piccinini & Craver’s (2011) claim that functional analyses can serve as “mechanism sketches”. The upshot lies within the free energy-principle’s unifying power: it can act as a kind of conceptual guide for revealing mechanistic explanations. Once physiological concepts are mapped onto the functional concepts derived from the free energy-principle, multilevel mechanistic explanations follow. But before this is elaborated the next section will give a short introduction to mechanistic explanation (Craver 2007).

4 Mechanistic explanation

Mechanistic explanation claims that in order “[t]o explain a phenomenon, [...] one has to

know what its components are, what they do and how they are organized [...]” (Craver & Kaplan 2011, p. 269). It does not suffice to merely be able, e.g. to accurately predict a phenomenon. Craver & Kaplan (2011, p. 271) show this by referring to the example of a heat gauge on a car. Despite the fact that the gauge represents engine heat and that one can also predict when the engine will overheat by looking at the gauge, it doesn’t explain why the engine is overheating. It only states that it is—not how it came about. Thus, mechanists introduced the “model-to-mechanism-mapping” (3M) requirement for explanatory models:

(3M) A model of a target phenomenon explains that phenomenon when (a) the variables in the model correspond to identifiable components and organizational features of the target mechanism that produces, maintains, or underlies the phenomenon, and (b) the causal relations posited among these variables in the model correspond to the activities or operations among the components of the target mechanism. (Kaplan 2011, p. 272)

This requirement can serve as a demarcation criterion as to when a model can actually be seen as explanatory. But how does mechanistic explanation progress? Two principal approaches are described by Craver & Kaplan (2011): reductionism and integrationism. The former tries to reduce mental phenomena into ever-smaller entities. Its most radical form, “ruthless reductionism”, is advocated by John Bickle (2003), who states that neuroscience should reduce “[...] psychological concepts and kinds to molecular-biological mechanisms and pathways” (Bickle 2006, p. 412). In other words, mental phenomena should be explained with low-level concepts. The integrationist approach, on the other hand, claims that explanations can be found across a hierarchy of mechanisms (Craver 2007), since every mechanism is itself embedded into a higher-level mechanism. Consequently, reductionism isn’t the only option, since “[...] mechanistic explanation requires consideration not just of the parts and operations in the mechanism but also of the organization

⁴ This does not mean that the free energy principle is false. On the contrary, this paper will present an attempt to increase its explanatory potential.

within the mechanism and the environment in which the mechanism is situated” (Bechtel 2009, p. 544). In particular, multilevel mechanistic explanations consider three viewpoints on any given mechanism: the etiological, constitutive, and contextual aspects (Craver 2013). At the etiological level, the causal history of a given mechanism is investigated at the same level of the hierarchy. Yet mechanisms can also be broken down into smaller, more specialised mechanisms. When investigating the internal mechanisms that give rise to a mechanism at a higher level, one can speak of the constitutive aspect of mechanistic explanation. This strategy resembles reductionism most. But, as mentioned before, every mechanism is also embedded in a higher-level mechanism. Thus, one must also investigate how a given mechanism contributes to the next higher-level mechanism. This has been termed the contextual aspect, because it situates a mechanism into a higher-order context. After this short introduction into mechanistic explanation, the next section will show how this relates to the problem above, i.e., that applications of the free energy principle operate with functional concepts and thus can’t serve as full explanations.

5 The free energy principle as heuristic guide

Here I will follow Piccinini & Craver’s (2011) proposal that functional descriptions are nothing other than mechanism sketches that derive their “[...] explanatory legitimacy from the idea that [they][...] capture something of the causal structure of the system” (Piccinini & Craver 2011, p. 306). Mechanism sketches are simply outlines of mechanisms that haven’t been fully investigated with regard to their structural properties. Thus, functional descriptions serve as placeholders until a mechanistic explanation can fully account for a given phenomenon by enriching functional concepts with concepts related to its structural properties.⁵ The explanatory gaps⁶ resulting from the functional nature of

the free energy principle could then be closed, leading to a shift from explanatory ambition to explanatory power. This also directly relates to the alleged preposterousness of the free energy principle, since the process of “filling-in” will diminish any residual doubts about the theory’s truthfulness. This can be applied to the free energy principle, which works with functional concepts such as “precision”, “prediction error”, “model optimization” or “attention”: “[o]nce the missing aspects are filled in, a functional analysis turns into a full-blown mechanistic explanation” (Piccinini & Craver 2011, p. 284). Take the concept of precision in the free energy principle as an example. As described above, precision gives an estimate concerning the “trustworthiness” of a given sensory signal and its ensuing prediction errors. Taken as such, precision is clearly a functional concept since it is “[...] specified in terms of effects on some medium or component under certain conditions” (Piccinini & Craver 2011, p. 291) without committing to any structural entities that could realise these functional properties. However, according to Friston et al. (2012), “[...] dopaminergic gating may represent a Bayes-optimal encoding of precision that enhances the processing of particular sensory representations by selectively biasing bottom-up sensory information (prediction errors)” (p. 2). In turn, “dopaminergic gating” involves the neurotransmitter dopamine, a molecule that can be structurally described. Crucially, now that the functional concept of precision, derived from the free energy principle, has been linked with dopaminergic gating, one can make further inferences as to how this entity is situated in a multilevel mechanism. For example, the modulation of precision has been associated with attention (Feldman & Friston 2010; Hohwy 2012), and since precision is realised via dopamine mediation, one can investigate the effects of dopamine on attentional mechanisms.⁷ On the other hand, if empirical evidence regarding precision or in particular predictions of precisions (hyperpriors) find “[...] that

⁵ However, as a preliminary note, both functional and structural properties are needed for a full mechanistic explanation (cf. Piccinini & Craver 2011, p. 290).

⁶ In this paper, the term “explanatory gap” is not used in the sense of “an explanatory gap [...] between the functions and experience”

(Chalmers 1995, p. 205; see Levine 1983 for the classical reference), as we see in the philosophy of mind. Rather, it describes the lack of neurobiological details in functional concepts.

⁷ Of course, to do so one would also have to know all the components involved in the mechanism responsible for attention.

descending signals do not mediate expected precisions, this would falsify the free energy principle” (p. 16). This further accentuates the need for mechanistic explanations.

As a more elaborate example, the phenomenon of biased competition will shortly be introduced. In biased competition, two stimuli are presented at a topographically identical location. However, only one of these stimuli is actually perceived. Thus the principal question: by which means does the brain “select” any given stimuli? In the free energy principle, the most obvious answer would be the stimulus that best minimises free energy or prediction error. However, in these cases, the stimuli are equally accurate, i.e., they both represent the causal structure of the world equally well. As a consequence, the stimuli will “[...] compete for the responses of cells in visual cortex” (Desimone 1998, p. 1245). Crucially, Desimone (1998) brings up a preliminary study by Reynolds et al. (1994) that states “[...] that attention serves to modulate the suppressive interaction between two or more stimuli within the receptive field [...]” (Desimone 1998, p. 1250). Thus, attention could be the determining factor as to which stimulus is perceived at a given moment. From the perspective of the free energy principle and in accordance with these findings, Feldman & Friston (2010) propose that “[...] attention is the process of optimizing synaptic gain to represent the precision of sensory information (prediction error) during hierarchical inference” (p. 2). These two views agree, since synaptic gain also entails a suppressive effect upon the other competing stimuli. Also, as just mentioned, Friston et al. (2012) identify precision weighting with dopaminergic gating, i.e., they argue that dopamine mediation realises the precision of incoming stimuli or prediction errors.

Now a fuller picture can be presented. This much more complete picture allows us to see how the free energy principle or prediction error minimization framework can prove to be beneficial with regard to mechanistic explanation. The phenomenon to be explained is biased competition. The mechanism that realises, or resolves, biased competition, i.e., the competition between two identically accurate and topo-

graphically identical stimuli, is precision weighting. This represents the etiological level of description since it describes how biased competition is resolved at a level of description that doesn’t refer to lower-level processes nor to how they are embedded into a higher order mechanism. It remains at the same level in the hierarchy of mechanisms. At the constitutive level we have the fact presented by Friston et al. (2012), that precision weighting is neurophysiologically realised by dopaminergic gating. This *constitutes* precision weighting and is located at a lower level. Last, precision weighting is embedded into the higher-order mechanism of attention. Precision weighting contributes to this higher order mechanism, or, from the other perspective, attention is constituted by precision weighting. This represents the contextual description.

The upshot is that, just as “[e]volutionary thinking can be heuristically useful as a guide to creative thinking about what an organism or organ is doing [...]” (Craver 2013, p. 20), the free energy principle can be a useful guide in finding multilevel mechanistic explanations concerning how the mind works. Due to its unifying power, the free energy principle offers a grand framework that seeks to explain every aspect of human cognition. Thus, filling increasingly more mechanistic concepts into functional placeholders will enable an understanding of the mind in terms of how it does work instead of how it ought to work. The explanatory worth of the free energy principle would then be preserved, since “[i]f these heuristics contribute to revealing some relevant aspects of the mechanisms that produce phenomena of interest, then Bayesian unification has genuine explanatory traction” (Colombo & Hartmann 2014, p. 3).

However, this should not be seen as an attempt to eliminate functional concepts by reducing them to mechanistic ones. Instead, as mentioned above, the integrationist account emphasises that functional and mechanistic concepts are both necessary for mechanistic explanations, since “structural descriptions constrain the space of plausible functional descriptions, and functional descriptions are elliptical mechanistic descriptions” (Piccinini & Craver 2011,

p. 307). Furthermore, once every functional term has a mechanistic counterpart, the 3M requirement posed by mechanists can be fulfilled in the case of the free energy principle.

Last, as a general remark, searching for structural properties seems important if researchers want to ground the free energy principle in the human brain. Functional theories are subject to multiple realizability. This means that not only humans or mammals could be bound to the free energy principle, but also Martians or bacteria or anything that could possess the “hardware” to do so. Hohwy suggests that the free energy principle can be seen as a biofunctionalist theory (this collection p. 20). In principle this means that the free energy principle can be multiply realised as long as that creature acts in such a way as to maintain itself in a certain set of expected states. These expected states then determine the creature’s phenotype. In seeking to explain human cognition, functional theories have to be enriched with mechanistic concepts relating to structural properties, since otherwise we could also be investigating Martians.

6 Conclusion

The negative thesis of this paper states that the free energy principle’s explanatory power, unlike its unificatory power, can be regarded as weak, since it does not fulfil the 3M requirement posited by mechanists. This follows from the fact that the free energy principle is a functional theory, thus also employing functional concepts. Yet these do not explain how a given phenomenon in fact does work but only how it should work. However, Piccinini & Craver (2011) propose that functional analyses, ultimately, are nothing else but mechanism sketches, i.e., incomplete mechanistic explanations.

In this paper I have tried to make a positive contribution to the discussion by arguing for two claims: first, since the free energy principle incorporates empirical results from psychology and neuroscience it provides a stronger case of unification (SUT) than the unification provided by Bayesian theory alone. By not solely relying

on its mathematical foundation, the free energy principle can try to ground its findings empirically in the brain. As a result, both the free energy principle and theories from psychology and neuroscience can constrain each other, thus being beneficiary to one another. Second, I argue that the free energy principle can act as a guide to finding multilevel mechanistic explanations. By linking mechanistic concepts with functional concepts from the free energy principle, the 3M requirement posited by mechanists can be fulfilled, consequently leading to actual explanations. This relates to the accused preposterousness of the free energy principle: with increasing explanatory power it becomes more and more difficult to deny that the free energy principle itself is, in fact, true.

References

- Bechtel, W. (2009). Looking down, around, and up: Mechanistic explanation in psychology. *Philosophical Psychology*, 22 (5), 543-564. [10.1080/09515080903238948](https://doi.org/10.1080/09515080903238948)
- Bickle, J. (2003). *Philosophy and neuroscience: A ruthlessly reductive account*. Dordrecht, NL: Kluwer Academic.
- (2006). Reducing mind to molecular pathways: Explicating the reductionism implicit in current cellular and molecular neuroscience. *Synthese*, 151, 411-434. [10.1007/s11229-006-9015-2](https://doi.org/10.1007/s11229-006-9015-2)
- Chalmers, D. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2 (3), 200-219. [10.1093/acprof:oso/9780195311105.003.0001](https://doi.org/10.1093/acprof:oso/9780195311105.003.0001)
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36 (3), 181-204. [10.1017/S0140525X12000477](https://doi.org/10.1017/S0140525X12000477)
- Colombo, M. & Hartmann, S. (2014). Bayesian cognitive science, unification, and explanation. [Pre-Print]. (Unpublished)
- Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. New York, NY: Oxford University Press.
- (2013). Functions and mechanisms: A perspectivalist view. *Synthese Library*, 363, 133-158. [10.1007/978-94-007-5304-4_8](https://doi.org/10.1007/978-94-007-5304-4_8)
- Craver, C. F. & Kaplan, D. M. (2011). Towards a mechanistic philosophy of neuroscience. In S. French & J. Saatsi (Eds.) *Continuum companion to the philosophy of science* (pp. 268-290). London, UK: Continuum Press.

- Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate visual cortex. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 353 (1373), 1245-1255. [10.1098/rstb.1998.0280](https://doi.org/10.1098/rstb.1998.0280)
- Feldman, H. & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, 4 (215), 1-23. [10.3389/fnhum.2010.00215](https://doi.org/10.3389/fnhum.2010.00215)
- Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Science*, 360 (1456), 815-836. [10.1098/rstb.2005.1622](https://doi.org/10.1098/rstb.2005.1622)
- (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11, 127-138. [10.1038/nrn2787](https://doi.org/10.1038/nrn2787)
- Friston, K. J., Daunizeau, J. & Kiebel, S. J. (2009). Active inference or reinforcement learning? *PLoS ONE*, 4 (7), e6421. [10.1371/journal.pone.0006421](https://doi.org/10.1371/journal.pone.0006421)
- Friston, K. J., Mattout, J. & Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics*, 104 (1-2), 137-160. [10.1007/s00422-011-0424-z](https://doi.org/10.1007/s00422-011-0424-z)
- Friston, K. J., Shiner, T., FitzGerald, T., Galea, J. M., Adams, R., Brown, H., Bestmann, S., Dolan, R. J., Moran, R. & Stephan, K. E. (2012). Dopamine, Affordance and Active Inference. *PLoS Computational Biology*, 8 (1), e1002327. [10.1371/journal.pcbi.1002327](https://doi.org/10.1371/journal.pcbi.1002327)
- Hesselmann, G., Sadaghiani, S., Friston, K. J. & Kleinschmidt, A. (2012). Predictive coding or evidence accumulation? False inference and neuronal fluctuations. *PLoS ONE*, 5 (3), e9926. [10.1371/journal.pone.0009926](https://doi.org/10.1371/journal.pone.0009926)
- Hobson, J. A. & Friston, K. J. (2012). Waking and dreaming consciousness: Neurobiological and functional considerations. *Progress in Neurobiology*, 98 (1), 82-98. [10.1016/j.pneurobio.2012.05.003](https://doi.org/10.1016/j.pneurobio.2012.05.003)
- Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology*, 3 (96), 1-14. [10.3389/fpsyg.2012.00096](https://doi.org/10.3389/fpsyg.2012.00096)
- (2013). *The predictive mind*. Oxford, UK: Oxford University Press.
- (2015). The neural organ explains the mind. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Jones, M. & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of bayesian models of cognition. *Behavioral and Brain Sciences*, 34 (4), 168-188. [10.1017/S0140525X10003134](https://doi.org/10.1017/S0140525X10003134)
- Kaplan, D. M. (2011). Explanation and description in computational neuroscience. *Synthese*, 183 (3), 339-373. [10.1007/s11229-011-9970-0](https://doi.org/10.1007/s11229-011-9970-0)
- Kaplan, D. M. & Craver, C. F. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy of Science*, 78 (4), 601-627. [10.1086/661755](https://doi.org/10.1086/661755)
- Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, 64, 354-361.
- Mumford, D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biological Cybernetics*, 66 (3), 241-251. [10.1007/BF00202389](https://doi.org/10.1007/BF00202389)
- Piccinini, G. & Craver, C. F. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, 183 (3), 283-311. [10.1007/s11229-011-9898-4](https://doi.org/10.1007/s11229-011-9898-4)
- Rao, R. P. N. & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2 (1), 79-87. [10.1038/4580](https://doi.org/10.1038/4580)

The Diversity of Bayesian Explanation

A Reply to Dominic L. Harkness

Jakob Hohwy

My claim is that, if we understand the function of the brain in terms of the free energy principle, then the brain can explain the mind. Harkness discusses some objections to this claim, and proposes a cautious way of solidifying the explanatory potential of the free energy principle. In this response, I sketch a wide, diverse, and yet pleasingly Bayesian conception of scientific explanation. According to this conception, the free energy principle is already richly explanatory.

Keywords

Bayesian explanation | Free-energy principle | Functionalism | Mechanistic explanation | Philosophy of neuroscience | Scientific explanation | Scientific unification

Author

Jakob Hohwy

jakob.hohwy@monash.edu

Monash University
Melbourne, Australia

Commentator

Dominic L. Harkness

dharkness@uni-osnabrueck.de

Universität Osnabrück
Osnabrück, Germany

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

The free energy principle (FEP) is ambitiously touted as a unified theory of the mind, which should be able to explain everything about our mental states and processes. Dominic L. Harkness discusses the route from the principle to actual explanations. He reasonably argues that it is not immediately obvious how explanations of actual phenomena can be extracted from the free energy principle, and then offers positive suggestions for understanding FEP's potential for fostering explanations. The argument I focus on in [Hohwy \(this collection\)](#) is that FEP is not so preposterous that it cannot explain at all; Harkness's com-

mentary thus raises the important point that there may be other obstacles to explanatoriness than being preposterous.

A further aspect of Harkness' approach is to make contact between the discussion of FEP's explanatory prowess and discussions in philosophy of neuroscience about computational and mechanistic explanation. This matters, since, if FEP is really set to dominate the sciences of the mind and the brain, then we need to understand it from the point of view of philosophy of science.

In this response, I will attempt to blur some distinctions between notions currently dis-

cussed in the philosophy of science. This serves to show that there is a diversity of ways in which a theory, such as FEP, can be explanatory. I am not, however, advocating explanatory pluralism; rather, I am roughly sketching a unitary Bayesian account of explanation according to which good explanation requires balancing the diverse ways in which evidence is explained away. This seems to me an attractive approach to scientific explanation—not least because it involves applying FEP to itself. The upshot is that even though FEP is not yet a full explanation of the mind, there are several ways in which it already now has impressive explanatory prowess.

2 Explanations, functions and mechanisms

Harkness employs existing views in the philosophy of science to create a divide between functions and mechanisms: functions specify what some phenomenon of interest ought to be doing, they don't specify how it actually does it. For that, a mechanism is needed which, in addition to specifying a functional role, also names the parts of the mechanism that perform this role (i.e., the realisers of the function), for example in the brain. This is thought to limit the explanatory power of FEP, which at its mathematical heart is just functionalist.

Whilst I accept the divide between functions and realisers, I don't think there is much explanatory mileage in naming realisers. If I already know what functional role is being realized, I don't come to understand a phenomenon better by being given the names of the realizing properties. This can be seen by imagining any mechanistic explanation (encompassing both functional role and realisers) where the names of the realizing properties are exchanged for other names. Such a move might deprive us of knowledge of which parts of the world realize this function, but this is not in itself explanatory knowledge. For example, I get to understand the heart by being told the functional role realized by atria and ventricles; I don't lose understanding if we rename the atria "As" and ventricles "Bs".

This is not to deny that we can gain understanding from learning about mechanisms. In particular, if I don't know about a phenomenon of interest, then I might explore the realizer of a particular case, and thereby get clues about the functional role. For example, in the 17th century William Harvey was able to finally comprehensively explain the functional role of the heart by performing vivisection on animals. Indeed, the point of such an exercise is to arrive at a clear and detailed description of a functional role (recall the difference between behaviourism and functionalism is that for the latter, the functional role is not just an input-output profile but also a description of the internal states and transitions between states).

Importantly, exploration (e.g., via vivisection, or via functional magnetic resonance imaging) of a mechanism is not the only way to eventually arrive at explanations. There can be multiple contexts of discovery. In particular, there can be very broad empirical observations as well as conceptual arguments. In the case of FEP, a key observation is that living organisms exist in this changing world. That is, organisms like us are able to maintain themselves in a limited number of states. This immediately puts constraints on any mechanistic explanation, which must cohere with this basic observation. Further, since an organism cannot know a priori what its expected states are, there must be an element of uncertainty reduction going on within the organism in order to estimate its expected states, or model. In a world with state-dependent uncertainty, this must happen through hierarchical inference. With these simple notions, FEP itself is well on its way to being established.

So I don't think it is explanatory power that is limited by being confined, as FEP fundamentally is, to functional roles. This mainly seems to impose a limit on our knowledge of *which* objects realize a given functional role, or it might curb our *progress* in finessing the functional role in question. Whereas it is right to say that FEP is limited because it is merely functional, this limit does not apply to its explanatory prowess.

3 Explanations and mechanism sketches

In assessing the explanatoriness of a functional theory like FEP it is useful, as Harkness proposes, to consider it as a mechanism sketch. Sketchiness, however, comes in degrees, and it is hard to think of any extant scientific account that is not sketchy in some respects—no matter how abundantly mechanistic it is. There doesn't seem to be any principled point at which a sketchy functional account passes over into being a non-sketchy mechanistic account. Rather, an account may become less and less sketchy as the full functional role and its realisers are increasingly revealed. This would be one respect in which the explanation in question would expand: more types and ranges of evidence would be explained, accompanied by a richer understanding of the functional workings of the mechanism.

The idea here is that mechanistic explanation comes in degrees, which makes it hard to say clearly when something is a mechanism sketch. Speaking of organs, consider again the case of the heart. Harvey is often said to have provided the first full account of pulmonary circulation, and it might be true that his account is less sketchy than that of his precursors, such as the much earlier Ibn al-Nafis. Yet even Harvey had areas of ignorance about the heart, and had to deduce some parts of his theory from his hypothesis about the overall function of the heart. Indeed, he readily acknowledges the difficulty of his project:

When I first gave my mind to vivisections, as a means of discovering the motions and uses of the heart, and sought to discover these from actual inspection, and not from the writings of others, I found the task so truly arduous, so full of difficulties, that I was almost tempted to think, with Fracastorius, that the motion of the heart was only to be comprehended by God. (Harvey 1889, p. 20)

A key question then is how sketchy FEP is—is it more like Harvey's rather comprehensive sketch of the heart, or is it like that of al-

Nafis? (If it is not completely misguided, like Galen's claim that there are invisible channels between the ventricles.) Harkness suggests that part of the attraction of FEP is that it comes with more empirical specification than mere Bayesian theory. It is true that much of the literature on FEP tries to map mathematical detail onto aspects of neurobiology. However, the mathematical detail of FEP itself is devoid of particular empirical fact—it is purely functionalist. (We might even say FEP is more fundamental than the Bayesian brain hypothesis, since the latter seems to be derivable from the former.)

However, this austerity with respect to specification of particular types of fact does not make FEP inherently sketchy. The starting point for FEP is the trivial but contingent fact that the world is a changing place and yet organisms exist—that is, that they can maintain themselves in a limited set of fluctuating states. This very quickly leads to the idea that organisms must be recapitulating (modelling) the structure of the world, and that they must be approximating Bayesian inference in their attempt to figure out what their expected states are.

This starting point for FEP gives us a lot of structure to look for in the brains of particular creatures. It calls for hierarchical structures the levels of which can encode sufficient statistics (means and variances) of probability distributions, pass these as messages throughout the system, and engage in explaining away and updating distributions over various time-scales. This has a much more mechanistic flavour than a more pure appeal to Bayes' rule, which leaves many more questions about the inferential mechanistics of the brain unanswered. (Part of the difference here is that FEP suggests that the brain implements approximate Bayesian inference, described in terms of variational Bayes.)

It is reasonable, then, to say that, even when stripped of extraneous neurobiological scaffolding, FEP is not inherently sketchy. It might not have the wealth of particular fact that would make it analogous to Harvey's theory of the heart. But it gives a surprisingly very

rich description of the functional role implemented by the brain of living organisms.

4 Explanation and types of functionalism

One might still insist on the point that Harkness raises, namely that, even if FEP is not particularly sketchy when stripped of empirical content, it is really only an account of what the system *should* do, rather than what it *actually* does. There is of course some truth to this, since the mathematical formulation of FEP is an idealization of a system engaged in variational Bayes.

However, perhaps FEP is in a peculiar functionalist category. Its starting point, as I mentioned earlier, is the trivial truth that organisms exist, from which it follows that they must be acting to maintain themselves in a limited set of states, from which it in turn follows that they must be reducing uncertainty about their model. Thus the function described by FEP is not about what the system should or ought to be doing but about what it *must* be doing, given the contingent fact that it exists.

This starting point differs from common-sense functionalism because it is not based on conceptual analysis but is instead based on a basic observation, plus statistical notions. It also differs from empirical functionalisms (cf. psychofunctionalism) because it does not specify functional roles in terms of proximal input–output profiles for particular creatures. Neither are the functional roles it sets out defined in terms of teleologically-defined proper functions (cf. teleosemantics), except in so far as it could be said that the proper function of an organism is to exist.

This category of functionalism, which I dubbed “biofunctionalism”, seems intriguingly different from other kinds of functionalism. It provides a foundational functional role, which *must* be realized in living organisms, and from which more specific processes can be derived (for perception, action, attention etc.). This differs from austere functionalisms, which only say how things ought to be working, and it differs from fully mechanistic functionalisms, which specify how particular types of things actually work.

5 Explanation by unification, and by mechanism revelation

Explanation in science is not just a matter of revealing the full detail of the parts and processes of mechanisms. Explanation is many things, as evidenced by the literature on the topic in philosophy of science. Most commonly, explanation is sought to reveal causes, and the contemporary discussion of mechanisms contributes substantially to this discussion. A different idea is that *unification* is explanatory—and yet explanation by unification is a multifaceted and disputed notion.

I think FEP explains by unification because it is a principle that increases our understanding of many very different phenomena, such as illusions, social cognition, the self, decision, movement, and so on (see *The Predictive Mind*, Hohwy 2013, for examples and discussion). FEP teaches us something new and unexpected about these phenomena, namely that they are all *related* as different *instances* of prediction-error minimization. For example, we are surprised to learn that visual attention and bodily movement are not only both engaged in prediction error minimization, they are essentially identical phenomena. FEP thus explains by providing a new, unified and coherent view of the mind.

In this manner, FEP is explanatory partly in ways that are separate from mechanistic explanation, and also from the discussion of how the functionalist and mechanistic approaches relate to each other.

6 Explanation is itself Bayesian

The comments I have provided so far appear to pull somewhat in different directions. I have argued that there is no sharp delineation between functional and mechanistic accounts, and yet I acknowledged that the functional aspects of FEP do set it apart from fully mechanistic accounts. I have argued that merely naming realisers is not explanatory, yet I have acknowledged that mechanistic accounts are explanatory. I have argued (with Harkness) that FEP explains by guiding particular mechanistic ac-

counts, but also by unification. In each of these cases, there seems to be much diversity, or even tension, in how FEP is said to be explanatory.

This diversity and tension, however, is by design. Explanation is not a one-dimensional affair; rather, a hypothesis, h , can be explanatory in a number of different ways. This can be seen by applying the overall Bayesian framework to scientific explanation itself. The strength of the case for h is consummate with how much of the evidence, e , h can explain away. As we know from the discussion of FEP, explaining away can happen in diverse ways: by changing the accuracy, the precision, or the complexity of h , or by intervening to obtain expected, high precision e . As discussed for FEP in Hohwy (this collection), we can also consider h 's ability to explain away e over shorter or longer time scales: if h has much fine-grained detail it will be able to explain away much of the short term variability in e but may not be useful in the longer term, whereas a more abstract h is unable to deal with fine-grained detail but can better accommodate longer prediction horizons.

Sometimes these diverse aspects of Bayesian explaining-away pull in different directions. For example, an attempt at unification via de-complexifying h may come at the loss of explaining some particular mechanistic instantiations. Conversely, an overly complex h may be overfitted and thereby explain away occurrent particular detail extremely well but be at a loss in terms of explaining many other parts of e .

In constructing a scientific explanation, how should one balance these different aspects of Bayesian explanation? Again we can appeal to FEP itself for inspiration: a good explanation minimizes prediction error on average and in the long run. That is, a good explanation should not generate excessively large prediction errors, and should be robust enough to persist successfully for a long time. This is intuitive, since we don't trust explanations that tend to generate large prediction errors, nor explanations that cease to apply once circumstances change slightly.

Formulating the goal of scientific explanation in this way immediately raises the question of what it means for prediction error to be

"large" or for a hypothesis to survive a "long time". The answer lies in expected precisions and context dependence. In building a theory, the scientist also needs to build up expectations for the precision (i.e., size) of prediction errors, and for the spatiotemporal structure of the phenomenon of interest. Not surprisingly, these aspects are also found in the conception of hierarchical Bayesian inference.

Achieving this balanced goal requires a golden-mean-type strategy: explanations should not be excessively general nor excessively particular, given context and expectations. That is, h should be able to explain away e in the long term without generating excessive prediction errors in the short term, as guided by expectations of precision and domain.

I think FEP is useful for attaining this golden mean, and that this is what makes FEP so attractive and promising. As a scientific hypothesis, it does not prioritise one type of explanatory aspect over another, but instead balances explanatory aspects against each other such that prediction error concerning the workings of the mind is very satisfyingly minimized on average and in the long run (and this indeed is the message of *The Predictive Mind*). Rather poetically, in my view, this means that we should evaluate FEP's explanatory prowess by applying it to itself.

7 Conclusion

I have agreed, to a large extent, with the points Harkness makes in his commentary. I have however also sought to suggest a more pluralistic perspective on scientific explanation. This ensures that the free energy principle, as it applies to the neural organ, has great potential to explain many aspects of the mind. I went one step further, however, and suggested that behind this explanatory pluralism lies a unified, Bayesian account of explanation, which perfectly mimics the unifying aspects of the free energy principle itself.

References

- Harvey, W. (1889). *On the motion of the heart and the blood in animals*. London, UK: George Bell & Sons.
- Hohwy, J. (2013). *The predictive mind*. Oxford, UK: Oxford University Press.
- (2015). The neural organ explains the mind. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.

Millikan's Teleosemantics and Communicative Agency

Pierre Jacob

Millikan's teleosemantic approach constitutes a powerful framework for explaining the continued reproduction and proliferation of intentional conventional linguistic signs, and thereby the stability of human verbal intentional communication. While this approach needs to be complemented by particular proximate psychological mechanisms, Millikan rejects the mentalistic psychological mechanisms, which are part of the Gricean tradition in pragmatics. The goal of this paper is to assess the balance between Millikan's teleosemantic framework and the particular proximate psychological mechanisms that she favors.

Keywords

Acceptance (compliance)/understanding | Communicative/informative intention | Conventions | Coordination | Direct/derived proper function | Etiological theory of functions | Imitative learning | Mindreading | Natural signs | Perception theory of verbal understanding

Author

Pierre Jacob
jacob@ehess.fr
Institute Jean Nicod
Paris, France

Commentator

Marius F. Jung
mjung02@students.uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Editors

Thomas Metzinger
metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Introduction

In this paper, I wish to revisit a topic that I addressed many years ago (cf. [Jacob 1997](#)) from a novel perspective. Much philosophy of mind of the latter part of the twentieth century has been devoted to naturalizing intentionality or the contents of mental representations. One of the landmarks of naturalistic philosophy of mind of the past thirty years is unquestionably Ruth Millikan's teleosemantic framework. Teleosemantic theories are teleological theories that seek to explain content by appealing to the *functions* of representations. Like most teleosemantic approaches, [Millikan \(1984, 2004\)](#) embraces an *etiological* conception of function, ac-

cording to which functions are *selected effects* ([Millikan 1984, 1989b; Neander 1991, 1995, 2004; Wright 1973](#)): the function of a trait is the effect caused by the trait that explains the continued reproduction (survival or proliferation) of past tokens of this trait.

Millikan's teleosemantic approach is particularly impressive for two related reasons. First, it applies in a single stroke to the contents of intentional *mental* representations, whose function is to mediate between pairs of cognitive mechanisms located within single brains, and also to the meanings of intentional *conventional* linguistic signs, whose function is

to mediate between pairs of cognitive mechanisms located in the brains of distinct individuals. Second, her overall teleological (or teleo-functional) approach, based on the etiological theory of functions, is meant to offer an account of the proliferation or continued reproduction of both biological entities and non-biological cultural things, such as linguistic and non-linguistic conventions.

Following Mayr (1961), evolutionary biologists and philosophers have long argued that the distinction between so-called *ultimate* and *proximate* explanations of biological traits (e.g., behaviors) is central to evolutionary theorizing. Roughly speaking, ultimate explanations address *why*-questions: for example, why do birds sing? Why does singing confer a selectional advantage (or greater fitness) to birds? Proximate explanations address *how*-questions: for example, what are the particular external circumstances which trigger singing in birds? What are the internal brain mechanisms that allow birds to sing?

The distinction between ultimate and proximate biological explanations raises some deep scientific and philosophical questions. One such question is whether ultimate explanations should be construed as non-causal answers to why-questions. Some philosophers have argued that ultimate explanations are *selectional* explanations based on natural selection. Natural selection can account for the prevalence of some trait in a population of individuals, but it cannot track the causal process whereby the trait is generated in each individual in the first place (Sober 1984, pp. 147–152; Dretske 1988, pp. 92–93; Dretske 1990, pp. 827–830). Other philosophers have replied that selectional explanations are causal explanations, on the grounds that no token of a trait whose type has been selected for fulfilling its (etiological) function could proliferate unless it was linked by a causal chain to the earlier production of the selected effect by ancestor tokens of the same type of trait (Millikan 1990, p. 808).¹

In this paper, I will not address such perplexing issues. I will simply accept the validity

of the distinction and assume that (whether ultimate explanations are causal explanations or not) ultimate and proximate explanations are complementary, not competing, explanations. Given that why-questions are fundamentally different from how-questions, it is likely that ultimate explanations offer few (if any) constraints on proximate explanations, and vice versa. I will further assume that the distinction carries over from biological to cultural evolution and applies to the evolution of human communication (cf. Scott-Philipps et al. 2011). In particular, Millikan's basic teleosemantic account of the proliferation of intentional conventional linguistic signs can usefully be construed as a kind of ultimate explanation of human (verbal and non-verbal) communication. Its main task is to address questions such as: what is the evolutionary or cultural function of human communication? Why do humans engage in communication at all? As with other kinds of ultimate explanations, it needs to be supplemented by specific proximate explanations whose role is to disclose the particular human cognitive capacities and mental processes whereby humans produce and understand intentional conventional signs.

The goal of this paper is to assess the balance between Millikan's broad teleosemantic approach to the cooperative function of human communication and the choice of particular proximate psychological mechanisms that she endorses. In particular, I will focus on her anti-mentalistic view, namely that verbal understanding of another's utterance is a kind of direct *perception* of whatever the utterance is about, and her correlative rejection of the basic Gricean pragmatic assumption that verbal understanding is an exercise in *mindreading*. One of the distinctive features of the human mindreading capacity is that it enables individuals to make sense of two kinds of agency: instrumental and communicative agency. In order to make sense of an agent's instrumental action, one must represent the contents of both her motivations and epistemic states. In order to make sense of an agent's communicative action, as Grice has basically argued, the addressee must infer what the agent is trying to convey, i.e., her communicative intention, whose very fulfilment

¹ For further discussion cf. Jacob 1997, pp. 256–269.

requires that it is recognized by the addressee. What is distinctive of human intentional communication is that it enables the communicative agent to cause her addressee to acquire new psychological states, and thereby to manipulate his mind.

Thus, I shall examine the contrast between the particular proximate mechanisms favored by Millikan and the Gricean pragmatic tradition. In the first section, I shall spell out the basic Gricean mentalistic framework. In the second section, I will spell out Millikan's teleosemantic machinery. In the third section, I will examine Millikan's view that verbal understanding is an extended form of perception. In the fourth section, I will examine the extent to which Millikan's account of conventions can support her rejection of the Gricean assumption that verbal understanding is an exercise in mindreading. Finally, in the last section, I will show that recent developmental findings in the investigation of early human social cognition are relevant to the controversy between Millikan and the Gricean tradition over the choice of proximate mechanisms underlying human communication.

2 The Gricean mentalistic picture of communicative agency

The Gricean mentalistic tradition rests on three basic related assumptions.²

- The first is the assumption that the complete process whereby an addressee contributes to the full success of a speaker's communicative act should be decomposed into two separable psychological sub-processes: a process of *understanding* (or *comprehension*) of the speaker's utterance and a process of *acceptance*, which in turn can be construed as the addressee's acquiring either a new belief or a new desire for action (depending on the direction of fit of the speaker's utterance). I'll call this the *separability* thesis.

² Although the relevance-based approach advocated by Sperber & Wilson (1986) and Wilson & Sperber (2004) departs in some interesting respects from Grice's (1969, 1989) own approach, I will nonetheless call their approach "Gricean" because, in the context of the present paper, the continuities between the two frameworks are far more important than the discontinuities.

- The Gricean mentalistic tradition also rests on the assumption that verbal understanding is an exercise in mindreading, whereby the addressee recognizes the complex psychological state that underlies the speaker's communicative act. I'll call this the *mindreading* thesis. (Clearly, the mindreading thesis is presupposed or entailed by the separability thesis.)
- Third, the Gricean mentalistic tradition further rests on a fundamental hypothesized asymmetry between what is required for understanding *instrumental* non-communicative agency and *communicative* agency. An agent intends her instrumental action to satisfy her desire in light of her belief, and the desirable outcome of her instrumental action can be recognized even if the agent fails to fulfil her goal or intention. But the intended effect of a speaker's communicative action, which is the addressee's understanding of what she means, cannot be achieved unless the speaker's intention to achieve this effect is recognized (cf. Sperber 2000, p. 130). Unlike purely instrumental agency, communicative agency is *ostensive* in the following sense. A speaker's communicative act is ostensive because its desirable outcome cannot be identified unless the addressee recognizes what the speaker intends to make manifest to him, i.e., what Sperber & Wilson (1986) call the speaker's *informative* intention. Thus, the Gricean tradition rests on the thesis of the *ostensive* nature of communicative agency (Sperber & Wilson 1986).

2.1 The mindreading thesis

On the picture of pragmatics which is part of the Gricean tradition of the past forty years broadly conceived, a human agent could not achieve a verbal or non-verbal act of intentional communication unless she had a complex psychological state, which Grice (1957) called the "speaker's meaning" and which he construed as a set of three interrelated intentions.³ First of all, by producing an utterance (or any other piece of ostens-

³ For brevity, I'll use "speaker" instead of "communicative agent". But of course not all communicative actions are verbal.

ive communicative behavior), the speaker must have the basic intention (i) to act on her addressee's mind, i.e., to cause him to acquire a new belief or a new desire (or intention) to perform some action. Second, the speaker must intend (ii) her addressee to recognize the content of her basic intention. Third, she must further intend (iii) her addressee's recognition of her basic intention (in accordance with (ii)) to play a major role in his fulfilling her basic intention (i).

In the following, I will adopt (Sperber & Wilson's (1986) simplified two-tiered account, according to which a communicative agent who produces an utterance has *two* (not three) inter-related intentions: an *informative* and a *communicative* intention, the first of which is nested within the other. She has the informative intention to make some state of affairs manifest to her addressee and also the further communicative intention to make her informative intention manifest to her addressee. So in this framework, the speaker's communicative intention is fulfilled by the addressee as soon as the latter recognizes (or understands) which state of affairs it is the speaker's informative intention to make manifest. But more is required for the speaker's informative intention to be fulfilled: the addressee must further accept the speaker's epistemic or practical authority. Depending on the direction of fit of the speaker's utterance, the addressee must either believe the fact which it is the speaker's informative intention to make manifest to him, or he must acquire the desire to act so as to turn into a fact the possible state of affairs which it is the speaker's informative intention to make manifest to him.

In a nutshell, much of (Sperber & Wilson's (1986) relevance-based framework rests on their insightful recognition that, on the broad Gricean picture of the speaker's meaning, the task of the addressee can be usefully divided into two basic psychological processes: one is the process whereby the addressee *understands* (or recognizes) the speaker's informative intention and the other is the process whereby he *fulfils* the speaker's informative intention. The first process involves the addressee's recognition of the speaker's informative intention, whereby the addressee fulfils the speaker's communicat-

ive intention that he recognize the speaker's informative intention. By recognizing the speaker's informative intention, the addressee comes automatically to both fulfil the speaker's communicative intention and to understand (or comprehend) the speaker's utterance. But for the addressee to *recognize* the speaker's informative intention is not *ipso facto* to *fulfil* it. So the second process needed for the success of the speaker's communicative act involves the addressee's fulfilment of the speaker's informative intention, whereby the addressee either accepts a new belief (in accordance with the content of the speaker's assertion) or forms a new desire to act (in accordance with the content of the speaker's request; cf. Jacob 2011).

2.2 The separability thesis

While the relevance-based account of communication clearly presupposes the mindreading thesis, Sperber (2001) has offered further support in favor of the separability thesis. Following Krebs & Dawkins (1984), Sperber (2001) has argued that for cooperative communication to stabilize in human evolution, it must be advantageous to both senders and receivers. Since the interests of speakers and hearers are not identical, the cooperation required for the stabilization of communication is vulnerable to deception. When her utterance is descriptive, the speaker can speak either *truthfully* or *untruthfully*. The addressee can either *trust* the speaker or not. The speaker is better off if her addressee trusts her and worse off if he distrusts her, whether or not the speaker is truthful. If the addressee trusts the speaker, then he is better off if the speaker is truthful and worse off if the speaker is not truthful, while the addressee remains unaffected if he distrusts the speaker.

Clearly, not every speaker is (or should be) granted equal epistemic or practical authority on any topic by every addressee. As Sperber et al. (2010) have further argued, given the risks of deception, it is likely that human cooperative communication would not have stabilized in human evolution unless humans had evolved mechanisms of *epistemic vigilance*, whereby they filter the reliability of descriptive utter-

ances. Focusing on a speaker's assertions at the expense of her requests, a speaker's epistemic authority depends to a large extent on the addressee's evaluation of her reliability (or trustworthiness), which in turn depends jointly on the addressee's evaluation of the speaker's competence on the topic at hand and on the addressee's representation of how benevolent are the speaker's intentions towards him. According to [Sperber et al. \(2010\)](#), an addressee's epistemic vigilance can apply to either or both the *source* of the information being communicated and its *content*.

3 Millikan's teleosemantic machinery

3.1 Teleosemantics and informational semantics

One of the first attempts at a naturalistic account of content (or intentionality) in the philosophy of mind was [Dretske's \(1981\)](#) informational semantics, according to which a sign or signal s carries information about property F iff there is a nomic (or lawful) covariation between instances of F and tokenings of s . As [Millikan \(1984, 2004\)](#) emphasized shortly after, informational semantics faces the puzzle of accounting for the possibility of *misrepresentation*. If the conditional probability that F is instantiated given s is 1, then how could s ever misrepresent instances of F ? This puzzle is neatly solved by teleological approaches: if a representation has a *function*, then it can fail to fulfil its function and thereby misrepresent what it is designed to represent ([Millikan 2004](#), Ch. 5). According to [Dretske's \(1988, 1995\)](#) own later attempt at preserving informational semantics as part of teleosemantics, a sign or signal s could not represent some property F unless s had the function of carrying information about (or indicating) instances of F .

[Millikan's \(1984, 2004\)](#) teleosemantic approach sharply departs from [Dretske's \(1988, 1995\)](#) information-based framework in at least two fundamental respects. First of all, in [Millikan's](#) earliest (1984) teleosemantic framework, there was no room for information-theoretic notions at all. In her later (1989a, 2004) work, she

argued that carrying information could not be a teleological function of a sign on the following grounds. Whether a sign carries information about some property depends on how the sign was *caused* or *produced*. But according to the etiological theory of functions, the function of a sign is one of its own *effects*, i.e., the selected effect that explains the continued reproduction of tokens of signs of this kind. How a sign was caused cannot be one of its effects, let alone its selected effect. If and when a sign happens to carry information about something, carrying information cannot be its selected effect, i.e., its etiological function.⁴

Second, [Dretske's \(1981\)](#) informational semantics could only be suitably naturalistic in the required sense if information is construed as the converse of nomological covariation (or necessity), i.e., as an entirely non-intentional and/or non-epistemic commodity. But as [Millikan \(2004, pp. 32–34\)](#) argues, if signal s could not carry information about F unless the probability that F is instantiated when s is tokened were 1 (in accordance with some natural law), then no animal could ever learn about F from perceiving tokens of s .

In her 2004 book, [Millikan](#) elaborates a notion of natural sign that is more “user-friendly” precisely because “it is at root an epistemic notion” ([Millikan 2004](#), p. 37). On [Millikan's \(2004\)](#) account, natural signs (e.g., tracks made by quail) are locally recurrent signs within highly restricted spatial and temporal domains: relative to one local domain, such tracks are natural signs of quail. Relative to a neighboring domain, the very same tracks are made by pheasants and are therefore natural signs of pheasants, not quail. Locally recurrent signs afford knowledge of the world for animals who can learn how to track the circumscribed domains relative to which they carry reliable information. Furthermore, locally recurrent natural signs can form transitive chains (or be productively embedded) within circumscribed do-

⁴ For significant discussion and defense of the view that it is the etiological function of mental representations to carry information, in response to Millikan's criticisms, cf. [Neander \(1995, 2007\)](#), [Godfrey-Smith \(2006\)](#) and [Shea \(2007\)](#). Cf. the recent exchange between [Neander \(2011\)](#) and [Millikan \(2011\)](#) For a criticism of Millikan's view, cf. [Pietroski \(1992\)](#) and see [Millikan's \(2000\)](#) reply.

mains. For example, retinal patterns can be a natural sign of tracks in the ground, which in turn may be a natural sign of quail within a circumscribed local domain. Perception is what enables non-human animals and humans alike to track the meanings of locally recurrent natural signs in their circumscribed domain of validity and thereby to acquire knowledge of the world (Millikan 2004, Ch. 4).

The application of Millikan's (1984, 2004) teleosemantic framework to the meanings of intentional *conventional* signs results from the combination of three related ingredients: (i) the etiological view of functions; (ii) acceptance of the sender-receiver structure as a necessary condition on the contents of intentional representations; and (iii) a naturalistic account of the reproduction of conventions.

3.2 The etiological conception of functions

As I said above, on the etiological view, the function of some trait is its selected effect that explains the continued reproduction of past tokens of this trait. This is what Millikan (1984) calls a device's *direct* proper function. But a device may also have what she calls a *derived* proper function. For example, it is the direct proper function of the mechanism of color change in the skin of chameleons to make them undetectable from the local background by predators. It is a derived proper function of this mechanism in a particular chameleon, Sam, at a particular place and time, to make the color of its skin match the color of its particular local background at that time so as to make it undetectable by predators there and then.

While Millikan's teleofunctional framework based on the etiological approach to functions primarily fits biological traits, it applies equally to non-biological items such as non-bodily tools—including public-language forms. For example, screwdrivers have the direct proper function of turning (driving or removing) screws. This is the effect of screwdrivers that explains their continued reproduction. Clearly, a screwdriver may also be intentionally used for the purpose of driving a screw with a particular metallic

structure, length, and diameter into a particular wooden material at a particular time and place. If so, then driving this particular screw into this particular wooden material at a particular time and place will be the derived proper function that this particular screwdriver inherits from the agent's intention.

3.3 The sender-receiver framework

According to the sender-receiver framework, a sign or signal *R* can be an *intentional* representation (as opposed to a *natural* sign) only if it is a *relatum* in a three-place relation involving two systems (or mechanisms), one of which is the sender (who produces *R*), the other of which is the receiver (who uses *R*). By application of the etiological view of functions, the sender (or producer) and the receiver (or consumer) have co-evolved so that what Millikan (1984, 2004) calls the *Normal* conditions for the performance of the function of one depends on the performance of the other's function and vice versa. In a nutshell, the producer and the consumer are *co-operative* devices, whose interests overlap and whose activities are beneficial to both. Thus, the cooperative ternary sender-receiver structure naturally applies to the contents of intentional mental representations that mediate between cognitive mechanisms located within a single organism.⁵

In virtue of the fact that intentional mental representations can have two basic directions of fit, the evolved cooperation between the producer and the consumer can take two basic forms. If and when the representation is *descriptive* or has a mind-to-world direction of fit, the producer's function is to make a sign *R*, whose content matches some state of affairs *S*, for the purpose of enabling (or helping) the consumer to perform its own task when and only when *S* obtains. If and when the representation is *directive* (or prescriptive) or has a world-to-mind direction of fit, the producer's function is to produce a representation whose content will guide the consumer's action, and it is the con-

⁵ Cf. Godfrey-Smith (2013) and Artiga (forthcoming) for further elaborate discussion of the requirement of cooperation as a condition on application of the sender-receiver structure.

sumer's function to make the world match the content of the sign by its own activities. Furthermore, Millikan (1995, 2004) has argued that the most primitive kinds of intentional mental representations (shared by humans and non-human animals) are what she calls *pushmi-pullyu* representations, which are at once descriptive and prescriptive, with both a mind-to-world and a world-to-mind direction of fit.

3.4 Conventional patterns

The third component of Millikan's teleosemantic approach to the meanings of intentional conventional signs involves her (1998) naturalistic account of *conventions*. On her account, so-called *natural conventionality* rests on two elementary characteristics: first, natural conventions are patterns that are *reproduced* (or that proliferate). Second, they are reproduced (or "handed down") "owing to precedent determined by historical accident, rather than owing to properties that make them more intrinsically serviceable than other forms would have been" (Millikan 2005, p. 188). The fact that conventions rest on historical precedent to a large extent accounts for their arbitrariness.⁶ On the basis of her naturalistic account of the continued reproduction of natural conventions, Millikan further offers a purportedly naturalistic account of the continued reproduction of conventional public-language signs, whose function is to coordinate the transfer of information between speakers and hearers. She thereby extends the cooperative ternary sender–receiver structure to the meanings of intentional conventional signs (or public-language forms) that mediate cognitive mechanisms located within pairs of distinct organisms.

Conventional public-language forms are tools or *memes* in Dawkins's (1976) sense: they have been selected and have accordingly been reproduced because they serve *coordinating* functions between a sender (the speaker) and a receiver (the addressee), whose interests overlap. But like any other tool, in addition to its direct *memetic* (or "stabilizing") function (which ex-

plains its continued reproduction), a particular token of some public-language form may also have a derived function or purpose, derived from the purpose of the speaker who produced it at a particular place and time. Thus, a token of a public-language form has two kinds of purposes: a memetic purpose and the speaker's purpose, which may or not coincide (cf. Millikan 1984, 2004, 2005).

4 Is verbal understanding an extended form of perception?

4.1 Perceiving the world through language

One basic problem raised by Millikan's account of the proliferation of intentional conventional signs is that one and the same linguistic form detached from its context of use may belong to different *memetic* families (or chains of reproductive events). In the reproductive process, what gets copied from one pair of sender-receivers to the next is not merely a linguistic form (e.g., "clear"), but the *use* of a linguistic form embedded in a particular *context*. This is why on Millikan's (2005, Ch. 10, section 3) view, the boundary between semantics and pragmatics is *blurry* and the process whereby a hearer tracks the memetic lineage of a conventional sign is a pragmatic process. On the teleosemantic approach, the hearer's task is to retrieve the appropriate context necessary for recognizing the correct memetic family (or lineage) to which a particular conventional sign belongs. In a nutshell, the hearers' task is to track the *domains* of intentional *conventional* signs.

Thus, it would appear that the hearer's task is quite similar to what is involved in tracking the restricted domain over which the information carried by a locally recurrent *natural* sign (e.g., tracks made either by quail or by pheasants) is valid. Since tracking the local domains over which the information carried by locally recurrent natural signs is a *perceptual* task, it is not surprising that Millikan has persistently urged that "in the most usual cases understanding speech is a form of direct perception of whatever speech is *about*. Interpreting

⁶ Including the arbitrariness of the relation between particular word-types and what they mean (sense and/or reference).

speech does not require making any inferences or having any beliefs about words, let alone about speaker intentions” (Millikan 1984, p. 62).⁷ Millikan (2004, p. 122) nicely illustrates her view that verbal understanding is an extended form of perception:

rain does not sound the same when heard falling on the roof, on earth, on snow, and on the water, even though it may be directly perceived as rain through any of these media. Exactly similarly, rain has a different sound when the medium of transmission is the English language (“It’s raining!”). And it sounds different again when the medium of transmission is French or German.

In a nutshell, “during Normal conversation, it is not language that is most directly perceived by the hearer but rather the world that is most directly perceived *through* language” (Millikan 2005, p. 207).

Furthermore, both ordinary and extended perception rest on *translation*, not inference: “the first steps in perception involve reacting to natural signs of features of the outer world by translating them into inner intentional representations of these outer features, for example, of edges, lines, angles of light sources in relation to the eye” (Millikan 2004, p. 118). In *normal* verbal communication, translation plays a two-fold role in mediating transfer from the speaker’s belief to the addressee’s belief. First, the speaker of a descriptive utterance translates her belief into a sentential conventional sign. Secondly, the addressee translates the content of the speaker’s utterance into his own new belief (Millikan 1984, 2004, 2005).

4.2 Ordinary and extended perception

Clearly, Millikan’s thesis that verbal understanding is an extended form of perception is not consistent with the Gricean thesis that verbal understanding is an exercise in *mindreading*. But on the face of it, the thesis

that verbal understanding is an extended form of perception (of whatever speech is *about*) itself is puzzling for at least three related reasons.⁸ First of all, as Millikan (2004, Ch. 9) herself recognizes, there is a major difference between the content of a perceptual representation of some state affairs and the verbal understanding of the content of another’s testimony about the very same state of affairs. At an appropriate distance and in good lighting conditions, one could not perceive a cup resting on a table without also perceiving its shape, size, color, texture, content, orientation, and spatial location with respect to the table, to any other object resting on the table, and especially to oneself. As Millikan (2004, p. 122) recognizes, unlike the content of testimony, the content of ordinary perception can be put at the service of action precisely because it provides information about the agent’s spatial relation to an object that is potentially relevant for action. But if an addressee located in a room next to the speaker’s room understands the content of the latter’s utterance of the sentence “There is a cup on the table”, he may endorse the belief that there is a cup on the table without having any definite expectation about the shape, size, color, texture, content, orientation, and spatial location of the cup with respect to himself, the table, or anything else.

Second, the thesis that verbal understanding is an extended form of perception ought to be restricted to the hearer’s verbal understanding of the meanings of *descriptive* utterances of indicative sentences with a mind-to-world direction-of-fit, which describe *facts* (or *actual* states of affairs). It cannot without further modifications be directly applied to the hearer’s verbal understanding of the meanings of *prescriptive* utterances of *imperative* sentences whereby a speaker *requests* an addressee to *act* so as to turn a *possible* (non-actual) state of affairs into a fact (or an actual state of affairs). Prescriptive utterances, which have a world-to-mind direction of fit, fail to describe any fact that could be directly perceived at all. So the question arises whether Millikan would be willing to en-

⁷ Cf. Millikan (2000, Ch. 6), Millikan (2004, Ch. 9), Millikan (2005, Ch. 10).

⁸ Cf. Recanati (2002) for a defense of Millikan’s thesis.

dorse the revised two-tiered thesis that (i) a verbal understanding of a speaker's descriptive utterance is the perception of whatever the utterance is about and (ii) a verbal understanding of a speaker's prescriptive utterance is to intend to perform whatever action is most likely to comply with the speaker's request.

Finally, testimony enables a speaker to convey beliefs whose contents far outstrip the perceptual capacities of either the speaker or her addressee. For example, an addressee may understand that the speaker intends to verbally convey to him her belief that there is no greatest integer, that democracy is the worst form of government except all those other forms that have been tried from time to time, or that religion is the opium of the people. But it does not make much sense to assume that either the speaker or her addressee could perceive what the speaker's utterance is about.

4.3 Tracking the domains of intentional conventional signs

Furthermore, the thesis that verbal understanding is an extended form of perception clearly rests on the assumption that the process whereby the hearer of a speaker's utterance tracks the memetic family of the intentional conventional sign used by the speaker is basically the same as the process whereby human and non-human animals track the meanings of locally recurrent natural signs in their circumscribed domain of validity. As I mentioned above, Millikan (2004) argues that perception is the basic process whereby animals track the meanings of locally recurrent natural signs in their circumscribed domain of validity. Crucially, one can track the meanings of locally recurrent natural signs within their circumscribed domain of validity *without* representing an agent's psychological state. So the question arises whether a hearer of a speaker's utterance could *always* track the memetic family of the intentional conventional signs used by a speaker *without* representing any of the speaker's psychological states.

In particular, as Recanati (2007) has argued, the question arises for descriptive utter-

ances containing at least four kinds of conventional expressions considered by Millikan (2004, Chs. 10–12): so-called unarticulated constituents in Perry's (1986) sense, incomplete definite descriptions, quantifiers, and possessives. Consider first an utterance of (1):

(1) It is raining.

It is unlikely that by an utterance of (1) a speaker means to assert that it is raining somewhere or other at the time of utterance. Instead, she is likely to mean that it is raining at the time of utterance and at the place of utterance (which remains unarticulated in the sentence). If by an utterance of (1), the speaker could *only* mean that it is raining at the place of utterance, then Millikan's claim that a hearer need not represent any of the speaker's psychological states for the purpose of tracking the local domains of intentional conventional signs might be vindicated. However, by an utterance of (1) on the phone, a speaker located in Paris may mean that it is raining in Chicago, not in Paris. Similarly, a French speaker located in Paris may use the incomplete description "the President" to refer, not to the French President, but instead to the President of the US.

For the purpose of understanding an utterance of a sentence containing a universal quantifier, as shown by example (2), the hearer must be able to properly restrict the domain of the quantifier:

(2) Everyone is asleep.

By an utterance of (2), the speaker presumably means to assert, not that everyone in the universe is asleep, but that everyone in some restricted domain (e.g., a relevant household) is asleep.⁹ The relevant restricted domain is the domain the speaker has in mind. Finally, by using the possessive construction "John's book", the speaker may have in mind many different relations between John and the book: she may mean the book written by John, the book

⁹ A nice example suggested by a referee is "There is no beer left", where the audience does not take the speaker to mean that there is no beer left in the universe, but instead in some properly restricted domain (e.g., some relevant fridge).

read by John, the book bought by John, the book sold by John, the book John likes, the book John dislikes, the book John just referred to in the conversation, the book John lost, the book John gave to the speaker, the book the speaker gave to John, the book the hearer gave to John, and so on. Unless the hearer hypothesizes what relation the speaker has in mind, he will fail to understand what the speaker means by her utterance of “John’s book”. In none of these four cases does it seem as if the hearer could recognize the memetic family of intentional *conventional* signs, i.e., track their relevant domains—unless he could represent the contents of some of the speaker’s *beliefs* or *assumptions*.

5 Conventions and belief-desire psychology

5.1 Teleosemantics and the separability thesis

Millikan’s thesis that verbal understanding is an extended form of perception is meant as an alternative to the Gricean thesis that verbal understanding is an exercise in mindreading. The further question arises to what extent Millikan’s teleosemantic account of the proliferation of public language conventions is consistent with the Gricean separability thesis, i.e., the distinction between verbal understanding and either acceptance (belief) or compliance. I will first argue that there is a restricted sense in which Millikan’s teleosemantics seems to be consistent with the separability thesis. But I will further argue that in a broader sense Millikan’s rejection of the mindreading thesis undermines the separability thesis.

On Millikan’s teleosemantic account, for a speaker’s descriptive utterance of an indicative sentence to meet the requirement of cooperation (and mutual interest) between the sender (or producer) and the receiver (or consumer), its direct proper function must be to cause the addressee to form a (true) belief. For a speaker’s *prescriptive* utterance of an imperative sentence to meet the requirement of cooperation, its direct proper function must be to cause the ad-

dresser to act in compliance with the content of the speaker’s request.

In the terminology of the relevance-based framework, a speaker who utters a descriptive utterance makes manifest to her addressee her communicative intention to make manifest her informative intention to make some fact manifest to him. The addressee may fulfil the speaker’s communicative intention by recognizing her informative intention and yet fail to fulfil her informative intention by resisting endorsing the relevant belief. A speaker who utters a prescriptive utterance makes manifest to her addressee her communicative intention to make manifest her informative intention to make manifest to him the desirability of turning some possible state of affairs into a fact by his own action. The addressee may fulfil the speaker’s communicative intention by recognizing her informative intention and yet fail to fulfil the speaker’s informative intention by resisting endorsing the intention to act in accordance with the speaker’s request.

Origgi & Sperber (2000, pp. 160–161), who subscribe to the Gricean thesis of the separability between verbal understanding and acceptance or compliance, have argued that the direct proper function of either a descriptive utterance or a prescriptive utterance could not be to reliably elicit the addressee’s response “at the level of belief or desire formation” (i.e., “the cognitive outputs of comprehension”), but instead “at an intermediate level in the process of comprehension”. Millikan might reply that according to her teleosemantic framework, an utterance may have a direct proper *function* and yet remain *unfulfilled*. If so, then the fact that an addressee may fulfil the speaker’s communicative intention (by recognizing her informative intention) and yet fail to fulfil the speaker’s informative intention seems entirely compatible with the teleosemantic framework.

However, to the extent that Millikan explicitly rejects the mindreading thesis, which is presupposed by the separability thesis, it is unlikely that she would find the separability thesis itself acceptable. On the relevance-based approach, it is a sufficient condition for securing what Austin (1975) called the “up-

take” (or success) of a communicative act (or speech act) that the speaker causes the addressee to fulfil the speaker’s communicative intention by recognizing her informative intention. It is not necessary that the addressee further fulfil the speaker’s informative intention. Successful communication does not require the addressee to accept either a new belief or a new desire, in accordance with the speaker’s informative intention. But on Millikan’s teleosemantic framework, failure of the addressee to comply with the speaker’s goal of causing the addressee to accept either a new belief or a new desire looks like a failure of the addressee to cooperate with the speaker’s conventional action, and therefore like a breakdown of the speaker’s communicative action. In fact, Millikan (2000, 2004, 2005) has offered two broad grounds for rejecting the mindreading thesis, both of which make it unlikely that she would support the separability thesis; the second of which is based on developmental evidence. I start with the non-developmental argument.

5.2 Cooperation and social conformity

First, Millikan (2004) rejects the mindreading thesis as part of her criticism of the reasoning that leads to the separability thesis: she rejects the joint assumptions that human predictions of others’ behavior are based on mindreading and that cooperation in human verbal communication is vulnerable to the risks of deception. On the one hand, she argues that “most aspects of social living involve cooperation in ways that benefit to everyone [...] for the most part, social cooperation benefits both or all parties. There is nothing mysterious about its evolution in this respect” (Millikan 2004, pp. 21–22). In a nutshell, Millikan argues that the urge to explain how the benefits of human communication are not offset by the risks of deception is misplaced on the grounds that the interests of speakers and hearers are sufficiently similar, if not identical.¹⁰

On the other hand, she argues that we use belief-desire psychology, not for prediction, but “for explanation after the fact” (Millikan 2004, p. 22). This is consonant with her (1984, pp. 67–69) earlier claim that while human adults have the ability to *reflect* on a speaker’s communicative intention if the automatic flow of conversation is interrupted for one reason or another, *normal* verbal understanding does *not* require representing the speaker’s communicative intention. Instead, *normal* verbal understanding should be construed as a *conventional* transfer of information whereby the speaker *translates* her belief into an utterance, whose meaning is in turn *translated* back by the addressee into a newly acquired belief.

Thus, Millikan rejects two of the major assumptions on which the separability thesis rests. She underestimates the gap between the interests of speakers and hearers in human communication and she minimizes the role of belief-desire psychology in the prediction of others’ behavior. Interestingly, her rejection of both assumptions rests in turn on her own competing account of communicative acts. As she puts it, “a surprise of this analysis of the conventional nature of the information-transferring function of the indicative is that believing what you hear said in the indicative turns out to be a conventional act, something one does in accordance with convention” (Millikan 2005, p. 46).¹¹ First of all, Millikan (2004, p. 23) argues that humans expect others to behave in conformity with social conventions, not on the basis of others’ beliefs and desires. Second, she further speculates that the conventional behaviors that are caused by a disposition to social conformity may derive from natural selection the memetic function of serving a coordinating function (*ibid.*).

Clearly, being disposed to social conformity and expecting others to be similarly disposed may help solve coordination problems (as shown by driving on one side of the road). However, being disposed towards social conformity is not sufficient to comply with social conventions. Compliance requires *learning*, i.e., the ac-

¹⁰ As Godfrey-Smith (2013, p. 45) observes, sameness of interests in human cooperation can be safely assumed in small contemporary communities, but not on a large scale, and nor in an evolutionary context.

¹¹ Note that this quote seems to presuppose the negation of the separability thesis.

quisition of relevant true *beliefs* about the contents of social conventions. Thus, the basic challenge for Millikan's claim that humans expect others to behave, not so much in accordance with the contents of their beliefs and desires, but in conformity with social conventions, is to offer an account of how humans come to learn and thereby know what it takes to act in conformity with social conventions.

5.3 Counterpart reproduction

This issue has been highlighted by the exchange between Tomasello's (2006) comments on Millikan's (2005) book and Millikan's (2006) response, which focuses on Millikan's (1998) thesis that many conventions, whose function it is to solve *coordination* problems, are reproduced by what she calls *counterpart* reproduction (or nuts and bolt reproduction). Typical coordination problems involve at least two partners, who share a common purpose that can be achieved only if each partner plays its assigned role, where both partners can be required to perform either the same act or two distinct complementary acts. In counterpart reproduction, when the respective roles of each partner require them to perform two different complementary acts, one typically adjusts her behavior to the other's and vice-versa. Counterpart reproduction is exemplified by, e.g., handshake reproduction, the reproduction of the respective postures assumed by men and women in traditional dancing, the reproduction of social distances appropriate for conversation, or the reproduction of the use of chopsticks for eating. Similarly, Millikan (2005, 2006) argues that counterpart reproduction also underlies the continued reproduction of conventional public-language signs.

Millikan (2005) further mentions open, partially or completely blind, conventional *leader-follower* co-ordinations involved in joint actions based on shared goals, whereby one agent (the leader) introduces a component of a pattern whose completion requires her partner (the follower) to perform a complementary component (*ibid.*, pp. 12–14). One example of open conventional leader-follower coordination is the

pattern whereby one agent selects her seat at an arbitrary table in a restaurant and her partner follows suit and selects his accordingly. One example of a partially blind conventional leader-follower coordination is the couch-moving pattern whereby the leader affords the follower anticipatory cues of her next move by ostensibly exaggerating her own movements, where the follower's familiarity with the pattern enables him to recognize the leader's ostensive cues and thereby to reproduce the complementary portion of the joint action. Another of Millikan's examples of a partially blind conventional leader-follower coordination is the US mailbox-flag convention, whereby the leader puts up a flag after she has placed mail in the mailbox and the postman picks up the mail after perceiving the flag.

Much comparative work by Tomasello and colleagues (reported by Tomasello et al. 2005 and summarized by Tomasello 2006, 2008) shows that while most communicative gestures in chimpanzees are learnt by ontogenetic ritualization, most communicative behaviors in human infants are acquired by imitative learning. As Tomasello (2006) argues, Millikan's own requirement that the reproduction of a conventional pattern depends on "the weight of precedent", not on its perceived intrinsic superior ability to produce a desired result, seems better fulfilled by a process of imitative learning than by a process of trial and error whereby one individual adjusts her behavior to another's. There seems to be nothing arbitrary (as there should if it were conventional) about an individual's adjusting her behavior to another's. While Tomasello (2006) does not deny that counterpart reproduction plays a significant role in cultural transmission, he disputes the claim that the output of counterpart reproduction qualifies as conventional.

Part of the gap between Millikan and Tomasello lies in what they take to be the proper unit for the analysis of the mechanism underlying the continued reproduction of conventional patterns involved in solving problems of coordination. While Tomasello focuses on the learning capacities of single individual minds, Millikan focuses on what can be achieved by the

reciprocal adjustments of pairs of cooperative partners. For example, when [Millikan \(2005, 2006\)](#) argues that counterpart reproduction underlies the continued proliferation of the custom of using chopsticks for eating in some cultures, she construes the convention of using chopsticks as a solution to the problem of coordination between pairs of partners, some of whom buy chopsticks and use them for eating and others who manufacture chopsticks. The latter would not manufacture chopsticks unless the former bought them and used them for eating. Conversely, the former would not buy them and use them for eating unless the latter manufactured them.

But of course, as Millikan is aware, this leaves open the question of how young children learn to use chopsticks for eating. As [Millikan \(2006, pp. 45–46\)](#) rightly observes, young human children understand their native language long before they can speak it. Nor can they learn to understand by imitating mature speakers: as she puts it, “they don’t watch how other people understand and then copy”. She further argues that young children would never understand their native tongue unless “their teachers” spoke to them, but “their teachers” would never speak to young children unless “they had had some reasonably successful experience” with previous listeners. This makes the continued reproduction of conventional public-language signs fit the pattern of counterpart reproduction. But still the question arises: how do young children learn to produce words of their native tongue? Vocal imitative learning may well play an important role (cf. [Hauser et al. 2002](#)). In a nutshell, according to Millikan the function of conventions is to solve coordination problems. She offers an elegant account of the proliferation of conventions based on counterpart reproduction. Her account must make room for the role of imitative learning in the way young human children learn either to use chopsticks for eating or to produce (and not just understand) words of their native tongues. As I shall argue in section 6.2, evidence shows that imitative learning in young children rests on their ability to construe the model’s demonstration as an ostensive communicative action. If so, then Mil-

likan’s view that counterpart reproduction underlies the proliferation of conventions must make room for the role of children’s ability to recognize the model’s communicative intention.

6 Teleosemantics and the puzzles of early human social cognition

6.1 Millikan’s developmental puzzle

To further undermine the mindreading thesis, [Millikan \(1984, 2000, 2004, 2005\)](#) has also appealed to findings from the developmental psychological investigation of early human social cognition, showing that “children younger than about four, although fairly proficient in the use of language, don’t yet have concepts of such things as beliefs, desires, and intentions” ([Millikan 2005, p. 204](#)). If such children do not have such concepts, then, unlike adults, they cannot reflectively engage in tasks of mindreading, i.e., in tracking the contents of others’ intentions, beliefs, and desires. To the extent that they can engage in verbal understanding, this further shows that verbal understanding cannot rest on mindreading (or belief–desire psychology).

As Millikan emphasizes, much developmental evidence shows that before they are at least four years old the majority of human children systematically *fail* elicited-response false-belief tasks. (In the terminology of developmental psychologists [Baillargeon et al. 2010](#), *elicited-response* tasks are tasks in which a participant is requested to generate an explicit answer in response to an explicit question.) For example, in the Sally-Anne test, after Sally places her toy in the basket, she leaves. While Sally is away, Anne moves Sally’s toy from the basket to the box. When Sally returns, participants, who know the toy’s actual location, are explicitly asked to predict where Sally (who falsely believes her toy to be in the basket) will look for her toy. The evidence shows that the majority of three-year-olds, “although quite proficient in the use of language” (in Millikan’s terms, [Millikan 2005, p. 204](#)), typically point to the box (i.e., the toy’s actual location), not to the basket where the agent falsely believes her toy to be (cf. [Wimmer & Perner 1983](#), [Baron-](#)

Cohen et al. 1985 and Wellman et al. 2001 for a meta-analysis).

Millikan assumes that the failure of most three-year-olds in such elicited-response false-belief tasks demonstrates that they lack what she calls a “representational theory of mind”. In a nutshell, she assumes that success at elicited-response false-belief tasks is a necessary condition for crediting an individual with a representational theory of mind (i.e., the ability to track the contents of others’ false beliefs). Acceptance of this assumption gives rise to Millikan’s developmental puzzle, which is “to understand how very young children can be aware of the intentions and of the focus of attention of those from whom they learn language without yet having this sort of sophisticated theory of mind” (Millikan 2005, p. 205). Before explaining why Millikan’s assumption is contentious, I shall briefly examine Millikan’s solution to her own puzzle.

Millikan’s solution involves three related ingredients, the most important of which is her thesis that normal verbal understanding is an extended form of perception (which does not require thinking about a speaker’s intention at all). Second, she argues that young children can understand the goal-directedness of a speaker’s communicative action without tracking the content of her communicative intention. Third, she argues that young children can understand the referential focus of a speaker’s attention without having a sophisticated theory of mind. As I understand it, much of the argument for the possibility of understanding the referential focus of a speaker’s attention without having a sophisticated theory of mind rests on the thesis that verbal understanding is an extended form of perception. As I have already expressed doubts about the thesis that verbal understanding is an extended form of perception, I shall now briefly examine the second thesis: that young children could understand the goal-directedness of a speaker’s communicative without tracking the content of her communicative intention.

Millikan (2005, pp. 206–207) offers two main reasons for granting young children the ability to recognize the goal-directedness of a speaker’s communicative action without grant-

ing them a full representational theory of mind. First, she argues that the evidence shows that mammals (dogs and cats and non-human primates, presumably, as well) lack a representational theory of mind but have the ability to recognize the goal-directedness of each other’s behavior. So by parity, very young children should also be granted the ability to recognize the goal-directedness of others’ actions, including speakers’ communicative actions. Second, she argues that communicative actions are cooperative actions. When young children are engaged in some cooperative action (including a communicative action) with a caretaker, they can easily keep track of the shared goal of the cooperative action, while tracking the focus of the speaker’s visual attention, without having a full representational theory of mind.

On the one hand, there is evidence that non-human primates recognize the goals of *instrumental* actions (Call & Tomasello 2008). On the other hand, there is also evidence that non-human primates—and birds as well—can discriminate *knowledgeable* agents (who know about, e.g., food from visual perception) from *ignorant* agents (who don’t know about food because their line of vision is obstructed) in *competitive* situations (Bugnyar 2011; Call & Tomasello 2008; Dally et al. 2006; Hare et al. 2001; Tomasello et al. 2003). But the question raised by Millikan’s puzzle is to understand what enables very young human children to make sense jointly of a speaker’s goal and the focus of her visual attention, when the speaker is performing a *communicative* action, not an *instrumental* action, in a *cooperative*, not a *competitive*, context. The fact that non-human primates can represent the goal of an agent’s *instrumental* action and discriminate a *knowledgeable* from an *ignorant* agent in a *competitive* context falls short of providing the required explanation.

Furthermore, two of Millikan’s assumptions are contentious in light of recent findings from developmental psychology. One is her assumption that young children could recognize the goal-directedness of speakers’ *communicative* actions without a representational theory of mind. The other is her assumption that success

at elicited-response false-belief tasks should be taken as a criterion for having the ability to track the contents of others' false beliefs (and therefore having a representational theory of mind). I shall start with the former, which amounts to denying the asymmetry between instrumental and communicative agency—which I earlier dubbed the thesis of the *ostensive* nature of human communicative agency.

6.2 The puzzle of imitative learning

The first relevant developmental finding, reported by [Gergely et al. \(2002\)](#), shows that approximately one-year-old human children (fourteen-month-olds) selectively imitate an agent's odd action. First, infants were provided with ostensive cues whereby an agent made manifest her intention to convey some valuable information by looking into the infants' eyes and addressing them in motherese. She then told the infants that she felt cold and covered her shoulders with a blanket. She finally performed an odd head-action whereby she turned a light box in front of her by applying her head, in two slightly different conditions. In the hands-occupied condition, she used her hands in order to hold the blanket around her shoulder while she executed the head-action. In the hands-free condition, she ostensibly placed her free hands on the table while she executed the head-action. [Gergely et al. \(2002\)](#) found that while 69% of the children replicated the head-action in the hands-free condition, only 21% did in the hands-occupied condition. In the hands-occupied condition, the majority of children used their own hands to turn the light box on. [Csibra & Gergely \(2005, 2006\)](#) further report that the asymmetry between infants' replication of the model's odd head-action in the hands-free and hands-occupied conditions vanishes if the model fails to provide infants with ostensive cues.

[Gergely & Csibra \(2003\)](#) have reported evidence that twelve-month-olds expect agents engaged in the execution of *instrumental* actions to select the most efficient action as a means towards achieving their goal (or goal-state), in the context of relevant situational constraints. So

the findings on imitation reported by [Gergely et al. \(2002\)](#) raise the following puzzle. Many more infants replicated the agent's head-action when the teleological relation between the agent's means and the agent's goal was opaque (in the hands-free condition) than when it was transparent (in the hands-occupied condition). Why did infants reproduce the agent's head-action more when it was a *less* efficient means of achieving the agent's goal of switching the light box on?

The Gricean thesis about the *ostensive* nature of communicative agency and the asymmetry between instrumental and communicative agency is relevant to answering this puzzle. Arguably, reception of ostensive signals prepared the infants to interpret the agent's action as a communicative, not an instrumental, action. It made manifest to the infants that the agent intended to make something novel and relevant manifest to them by her subsequent non-verbal communicative action. In the hands-occupied condition, the infants learnt how contact was necessary in order to turn on the light bulb, which was part of an unfamiliar device. Since the model's hands were occupied, the infants whose own hands were free assumed that that they were free to select the most efficient means at their disposal to achieve the same goal as the model. In the hands-free condition, the model could have used her hands, but she did not. So the infants learnt from the model's non-verbal demonstration that they could turn the light on by applying their own heads.

On the one hand, the evidence shows that infants construe imitative learning as a response to an agent's communicative action and that they selectively imitate a model's action as a function of what they take to be relevantly highlighted by the model's communicative act (cf. [Southgate et al. 2009](#)). On the other hand, further evidence shows that newborns prefer to look at faces with direct gaze over faces with averted gaze. Right after birth, they display sensitivity to eye-contact, infant-directed speech or motherese, and infant-contingent distal responsiveness. If preceded by ostensive signals, an agent's gaze shift has been shown to generate in preverbal human infants a referential expecta-

tion, i.e., the expectation that the agent will refer to some object (Csibra & Volein 2008, cf. Csibra & Gergely 2009, and Gergely & Jacob 2013, for review).

One further intriguing piece of evidence for the early sensitivity of human toddlers to the ostensive nature of human communicative agency is offered by experiments that shed new light on the classical A-not-B perseveration error phenomenon first reported by Piaget (1954). Infants between eight and twelve months are engaged in an episodic hide-and-seek game in which an adult repeatedly hides a toy under one (A) of two opaque containers (A and B) in full view of the infant. After each hiding event, the infant is allowed to retrieve the object. During test trials where the demonstrator places the object repeatedly under container B, infants continue to perseveratively search for it under container A where it had been previously hidden. Experimental findings reported by Topal et al. (2008) show that minimizing the presence of ostensive cues results in significant decreases of the perseverative bias in ten-month-olds. This finding is consistent with the assumption that infants do not interpret the hide-and-seek game as a game, but instead as a teaching session about the proper location of a toy.

All this evidence strongly suggests that human infants are prepared from the start to recognize nonverbal ostensive referential signals and action-demonstrations addressed to them as encoding an agent's communicative intention to make manifest her informative intention to make some relevant state affairs manifest to the addressee. But of course this raises a puzzle: how could preverbal infants recognize an agent's communicative intention to make manifest her informative intention? A novel approach to this puzzle has been insightfully suggested by Csibra (2010). According to Csibra, very young infants might well be in a position similar to that of a foreign addressee of a verbal communicative act, who is unable to retrieve a speaker's informative intention for lack of understanding of the meaning of the speaker's utterance. Nonetheless, the foreign addressee may well recognize being the target of the speaker's communicative intention on the basis of the speaker's ostensive

behavior. Furthermore, ostensive signals to which preverbal human infants have been shown to be uniquely sensitive can plausibly be said to *code* the presence of an agent's communicative intention. If this is correct, then little (if any) further work would be left for preverbal infants to infer the presence of a speaker's communicative intention after receiving ostensive signals.

6.3 The puzzle about early false-belief understanding

As Millikan has emphasized, much developmental psychology has shown that the majority of three-year-olds fail *elicited-response* false-belief tasks. For example, when asked to predict where an agent with a false belief will look for her toy, most three-year-olds who know the toy's location point to the toy's actual location, and not to the empty location where the mistaken agent believes her toy to be. However, in the past ten years or so, developmental psychologists have further designed various *spontaneous-response* false-belief tasks, in which participants are *not* asked any question and therefore *not* requested to produce any answer. Typical spontaneous-response tasks involve the use of the violation-of-expectation and anticipatory-looking paradigms, which involve two steps. In habituation or familiarization trials, participants are first experimentally induced to form expectations by being repeatedly exposed to one and the same event. Second, in test trials of violation-of-expectation experiments, participants are presented with either an expected or an unexpected event. By measuring the time during which participants respectively look at the expected vs. the unexpected event, psychologists get evidence about the nature and content of the infants' expectations formed during the habituation or familiarization trials. Psychologists can also use the anticipatory-looking paradigm and experimentally determine where participants first look in anticipation of the agent's action, thereby revealing their expectation about the content of the agent's belief.

Thus, in a seminal study based on the violation-of-expectation paradigm by Onishi & Baillargeon (2005), fifteen-month-olds saw an

agent reach for her toy either in a green box or in a yellow box when she had either a true or a false belief about her toy's location. Onishi and Baillargeon report that fifteen-month-olds looked reliably longer when the agent's action was incongruent rather than congruent with the content of either her true or false belief. In a study based on the anticipatory looking paradigm, twenty-five-month-olds were shown to look correctly towards the empty location where a mistaken agent believed her toy to be, in anticipation of her action (Southgate et al. 2007). Many further subsequent studies show that toddlers and even preverbal human infants are able to track the contents of others' false beliefs and expect others to act in accordance with the contents of their true and false beliefs.

In a classical experiment by Woodward (1998), six-month-olds were familiarized with an agent's action, who repeatedly chose one of two toys. In the test trials, the spatial locations of the toys were switched and the infants either saw the agent select the same toy as before at a new location or a new toy at the old location. six-month-olds looked reliably longer at the former than at the latter condition. Luo & Baillargeon (2005) further showed that infants do not look reliably longer at a change of target if, in the familiarization trials, the agent repeatedly reached for the same object, but there was no competing object (for further discussion cf. Jacob 2012). This result has been widely interpreted as showing that six-month-olds are able to ascribe a preference to an agent. Luo (2011) further found that ten-month-olds who know that an agent is in fact confronted with only *one* object (not two) ascribe a preference to the agent if she *falsely believes* that she is confronted with a pair of objects, but *not* if the agent knows (as the infants do) that she is confronted with only one object.

Thus, the psychological investigation of early human social cognition is currently confronted with a puzzle different from that confronted by Millikan: on the one hand, robust findings show that the majority of three-year-olds fail elicited-response false-belief tasks such as the Sally-Anne test. On the other hand, more recent findings based on spontaneous-response

tasks show that preverbal infants expect others to act in accordance with the contents of their true and false beliefs. The puzzle is: how do we make sense of the discrepancy between both sets of experimental findings?

So far, psychologists have offered two broad strategies for this, one of which assumes (as Millikan does) that success at elicited-response false-belief tasks is a necessary condition of the ability to ascribe false beliefs to others, which is taken to be the output of "a cultural process tied to language acquisition" (Perner & Ruffman 2005, p. 214). Their burden is to explain away the findings about preverbal infants without crediting them with the ability to track the contents of others' false beliefs. Thus, the majority of "cultural constructivist" psychologists offer low-level associationist accounts of the findings about preverbal infants based on spontaneous-response tasks. Other psychologists (including Baillargeon et al. 2010; Bloom & German 2000 Leslie 2005; Leslie et al. 2004; Leslie et al. 2005; Scott et al. 2010) argue that the findings about preverbal infants show that they can track the contents of others' false beliefs. Their burden is to explain why elicited-response false-belief tasks are so challenging for three-year-olds. The prevalent non-constructivist explanation is the processing-load account offered by Baillargeon and colleagues.

The core of the associationist strategy is to account for findings about preverbal human infants based on spontaneous-response tasks on the basis of a three-way association between the agent, the object, and its location. It postulates that infants will look longer in the test trials at events that depart more strongly from the three-way association generated by the familiarization trials. For example, in the test trials of Onishi & Baillargeon (2005), infants should look longer when the agent reaches for her toy in the yellow box if in the familiarization trials the agent placed her toy in the green box on three repeated occasions.

The main obstacle for the associationist path is a recent study by Senju et al. (2011) based on the anticipatory-looking paradigm. In the familiarization stage, eighteen-month-olds experience the effect of wearing either an

opaque blindfold through which they cannot see or a *trick* blindfold through which they can see. In the first trials of the test phase, the children are familiarized to seeing an agent retrieve her toy at the location where a puppet has placed it in front of her. The agent's action is always preceded by a pair of visual and auditory cues. In the last test trial, the agent first sees the puppet place the toy in one of the two boxes; she then ostensibly covers her eyes with a blindfold, and finally the puppet removes the toy. After the puppet disappears, the agent removes her blindfold and the cues are produced. Using an eye-tracker, [Senju et al. \(2011\)](#) found that only infants who had experienced an opaque blindfold, not infants who had experienced a trick see-through blindfold, reliably made their first saccade towards the empty location in anticipation of the agent's action.

[Senju et al.'s \(2011\)](#) findings are inconsistent with the associationist strategy: since all infants saw exactly the same events, they should have formed exactly the same threefold association between the agent, the toy, and the location, and on this basis they should have gazed at the same location in anticipation of the agent's action. But they did not. Only infants whose view had been previously obstructed by an opaque blindfold, not those whose view had not been obstructed by a trick blindfold, expected the blindfolded agent to mistakenly believe that the object was still in the opaque container after the puppet removed it.

The evidence against the associationist strategy is also evidence against the assumption (accepted by Millikan) that success at elicited-response false-belief tasks is a necessary condition for having a representational theory of mind and being able to track the contents of others' false beliefs. But this assumption is unlikely to be correct if, as several critics of the cultural constructivist strategy have argued, the ability to ascribe false beliefs to others is not a sufficient condition for success at elicited-response false-belief tasks. As advocates of the processing-load account ([Baillargeon et al. 2010](#)) have argued, an agent could have the ability to ascribe false beliefs to others and still fail elicited-response false-belief tasks for at

least three reasons: she could fail to understand the meaning of the linguistically-encoded sentence used by the experimenter to ask the question. She could fail to select the content of the agent's false belief in the process whereby she answers the experimenter's question. She could fail to have the executive-control resources necessary to inhibit the prepotent tendency to answer the question on the basis of the content of her own true belief. I will now argue that solving the puzzle about early belief-understanding may well depend on acceptance of the Gricean thesis of the ostensive nature of communicative agency and the asymmetry between instrumental and communicative agency.

I now want to offer a speculative solution to the puzzle about early false-belief understanding based on two related Gricean assumptions. The first is the asymmetry between the non-ostensive nature of instrumental agency and the ostensive nature of human communicative agency. The second related assumption is that the human ability to track the content of the false belief of an agent of an instrumental action must be a by-product of the ability to deal with deception (e.g., lying) in the context of human communicative agency.

In the typical Sally-Anne elicited-response false-belief task, participants are requested to make sense of two actions performed by two different agents at the same time: they must track the contents of the motivations and epistemic states of a mistaken agent engaged in the execution of an instrumental action (Sally) and they must also make sense of the communicative action performed by the experimenter who asks them "Where will Sally look for her toy?" The findings based on spontaneous-response tasks strongly suggest that much before they become proficient in language use, young human children are able to spontaneously track the contents of the false beliefs of agents of instrumental actions. So the question is: what is it about the experimenter's question that biases them towards pointing to the toy's actual location?

In [Helming et al. \(2014\)](#), we have argued that two biases are at work, one of which is a referential bias and the other of which is a co-

operative bias. The referential bias itself turns on two components. On the one hand, the experimenter could not ask the question “Where will the agent look for her toy?” unless she referred to the toy. On the other hand, the experimenter shares the participants’ correct epistemic perspective on the toy’s location. In answering the experimenter’s question, participants have the option of mentally representing either the toy’s actual location or the empty location (where the mistaken agent believes her toy to be). The experimenter’s question may bias young children’s answer towards the actual location by virtue of the fact that the experimenter both referred to the toy (whose actual location they know) and shared the participants’ correct epistemic perspective on the toy’s actual location (at the expense of the mistaken agent’s incorrect perspective on the empty location). What we further call the co-operative bias is the propensity of young children to help an agent with a false belief about her toy’s location achieve the goal of her instrumental action by pointing to the actual location (cf. Warneken & Tomasello 2006, 2007; Knudsen & Liszkowski 2012), in accordance with their own true belief about the toy’s actual location. If so, then young children might interpret the prediction question “Where *will* Sally look for her toy?” as a normative question: “Where *should* Sally look for her toy?” Of course, the correct answer to the normative question is the toy’s actual location, not the empty location where the mistaken agent believes her toy to be.

7 Conclusion

The goal of this paper was to assess the gap between Millikan’s particular views about some of the proximate psychological mechanisms underlying human communication and three core assumptions of the Gricean approach: the mindreading thesis, the separability thesis, and the ostensive nature of communicative agency. I have criticized five of Millikan’s basic claims about psychological mechanisms: (i) verbal understanding is best construed as an extended form of perception; (ii) hearers can track the domains of intentional conventional signs

without representing any of the speaker’s psychological states; (iii) the overlap between the interests of speakers and hearers undermines the separability thesis; (iv) humans can predict others’ behavior out of social conformity; (v) developmental psychology supports the view that neither verbal understanding nor language acquisition requires a representational theory of mind.

Millikan’s major teleosemantic contribution has been to open an entirely novel approach to the continued reproduction of intentional conventional public-language signs. As was shown by the discussion of whether her view of the proper function of descriptive and prescriptive utterances is consistent with the separability thesis, there is room for disagreement about particular psychological mechanisms within a teleosemantic approach. I do not think that Millikan’s teleosemantic framework for addressing the continued reproduction of intentional conventional signs mandates the particular choice of proximate psychological mechanisms that she recommends. One of the major challenges for the scientific investigation of cultural evolution is to make sure that the proximate psychological mechanisms that underlie the continued reproduction of human cultural conventions are supported by findings from experimental psychological research, in particular developmental psychology.

Acknowledgements

I am grateful to the editors for inviting me to write this essay, and to Ned Block, Carsten Hansen, Georges Rey, Dan Sperber and two anonymous reviewers for their comments.

References

- Artiga, M. (forthcoming). *Signaling without cooperation*.
- Austin, J. (1975). *How to do things with words*. Oxford, UK: Oxford University Press.
- Baillargeon, R., Scott, R. M. & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences*, 14 (3), 110-118. [10.1016/j.tics.2009.12.006](https://doi.org/10.1016/j.tics.2009.12.006)
- Baron-Cohen, S., Leslie, A. & Frith, U. (1985). Does the autistic child have a “theory of mind?”. *Cognition*, 21 (1), 37-46. [10.1016/0010-0277\(85\)90022-8](https://doi.org/10.1016/0010-0277(85)90022-8)
- Bloom, P. & German, T. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, 77 (1), 25-31. [10.1016/S0010-0277\(00\)00096-2](https://doi.org/10.1016/S0010-0277(00)00096-2)
- Bugnyar, T. (2011). Knower-guesser differentiation in ravens: others’ view points matter. *Proceedings of the Royal Society of London B: Biological Sciences*, 278 (1705), 634-640. [10.1098/rspb.2010.1514](https://doi.org/10.1098/rspb.2010.1514)
- Call, J. & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, 12 (5), 187-192. [10.1016/j.tics.2008.02.010](https://doi.org/10.1016/j.tics.2008.02.010)
- Csibra, G. (2010). Recognizing communicative intentions in infancy. *Mind and Language*, 25 (2), 141-168. [10.1111/j.1468-0017.2009.01384.x](https://doi.org/10.1111/j.1468-0017.2009.01384.x)
- Csibra, G. & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, 13 (4), 148-153. [10.1016/j.tics.2009.01.00](https://doi.org/10.1016/j.tics.2009.01.00)
- Csibra, G. & Volein, A. (2008). Infants can infer the presence of hidden objects from referential gaze information. *British Journal of Developmental Psychology*, 26 (1), 1-11. [10.1348/026151007X185987](https://doi.org/10.1348/026151007X185987)
- Dally, J. M., Emery, N. J. & Clayton, N. S. (2006). Food-caching western scrub-jays keep track of who was watching when. *Science*, 312 (5780), 1662-1665. [10.1126/science.1126539](https://doi.org/10.1126/science.1126539)
- Dawkins, R. (1976). *The selfish gene*. Oxford, UK: Oxford University Press.
- Dretske, F. (1981). *Knowledge and the flow of information*. Cambridge, MA: MIT Press.
- (1988). *Explaining behavior*. Cambridge, MA: MIT Press.
- (1990). Reply to reviewers of explaining behavior: Reasons in a world of causes. *Philosophy and Phenomenological Research*, 50 (4), 819-839.
- (1995). *Naturalizing the mind*. Cambridge, MA: MIT Press.
- Gergely, G. & Csibra, G. (2003). Teleological reasoning in infancy: The naive theory of rational action. *Trends in Cognitive Sciences*, 7 (7), 287-292. [10.1016/S1364-6613\(03\)00128-1](https://doi.org/10.1016/S1364-6613(03)00128-1)
- (2005). The social construction of the cultural mind: imitative learning as a mechanism of human pedagogy. *Interaction Studies*, 6 (3), 463-481. [10.1075/bct.4.19ger](https://doi.org/10.1075/bct.4.19ger)
- (2006). Sylvia’s recipe: the role of imitation and pedagogy in the transmission of cultural knowledge. In S. Levenson & N. Enfield (Eds.) *Roots of human sociality: Culture, cognition, and human interaction* (pp. 229-255). Oxford, UK: Berg Publishers.
- Gergely, G., Bekkering, H. & Kiraly, I. (2002). Rational imitation in preverbal infants. *Nature*, 415 (6873), 755-755. [10.1038/415755a](https://doi.org/10.1038/415755a)
- Gergely, G. & Jacob, P. (2013). Reasoning about instrumental and communicative agency in human infancy. In J. B. Benson, F. Xu & T. Kushnir (Eds.) *Rational constructivism in cognitive development* (pp. 59-94). Waltham, MA: Academic Press.
- Godfrey-Smith, P. (2006). Mental representation, naturalism, and teleosemantics. In G. MacDonald & D. Papineau (Eds.) *Teleosemantics* (pp. 42-68). Oxford, UK: Oxford University Press.
- (2013). Signals, icons, and beliefs. In D. Ryder, J. Kinsbury & K. Williford (Eds.) *Millikan and her critics* (pp. 41-58). Oxford, UK: Blackwell.
- Grice, H. P. (1957). Meaning. *The Philosophical Review*, 66 (3), 377-388. [10.2307/2182440](https://doi.org/10.2307/2182440)
- (1969). Utterer’s meaning and intentions. *The Philosophical Review*, 78 (2), 147-177. [10.2307/2184179](https://doi.org/10.2307/2184179)
- (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.
- Hare, B., Call, J. & Tomasello, M. (2001). Do chimpanzees know what conspecifics know? *Animal Behaviour*, 61 (1), 139-151. [10.1006/anbe.2000.1518](https://doi.org/10.1006/anbe.2000.1518)
- Hauser, M., Chomsky, N. & Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science*, 298 (5598), 1569-1579. [10.1126/science.298.5598.1569](https://doi.org/10.1126/science.298.5598.1569)
- Helming, K.A., Strickland, B. & Jacob, P. (2014). Making sense of early false-belief understanding. *Trends in Cognitive Sciences*, 18 (4), 167-170. [10.1016/j.tics.2014.01.005](https://doi.org/10.1016/j.tics.2014.01.005)
- Jacob, P. (1997). *What minds can do*. Cambridge, UK: Cambridge University Press.
- (2011). Meaning, intentionality and communication. In C. Maierborn, K. Heusinger & P. Portner (Eds.) *Semantics: an international handbook of natural language meaning* (pp. 11-24). Berlin, GER: Walter de

- Gruyter.
- (2012). Sharing and ascribing goals. *Mind and Language*, 27 (2), 200-227. [10.1111/j.1468-0017.2012.01441.x](https://doi.org/10.1111/j.1468-0017.2012.01441.x)
- Knudsen, B. & Liszkowski, U. (2012). 18-month-olds predict specific action mistakes through attribution of false belief, not ignorance, and intervene accordingly. *Infancy*, 17, 672-691.
- Krebs, J. R. & Dawkins, R. (1984). Animal signals: Mind-reading and manipulation. In J. R. Krebs & N. B. Davies (Eds.) *Behavioral ecology* (pp. 380-402). Sunderland, MA: Sinauer Associates.
- Leslie, A. (2005). Developmental parallels in understanding minds and bodies. *Trends in Cognitive Sciences*, 9, 459-462.
- Leslie, A. M., Friedman, O. & German, T. P. (2004). Core mechanisms in “theory of mind”. *Trends in Cognitive Sciences*, 8 (12), 528-533.
- Leslie, A. M., German, T. P. & Polizzi, P. (2005). Belief-desire reasoning as a process of selection. *Cognitive Psychology*, 50, 45-85.
- Luo, Y. (2011). Do 10-month-old infants understand others’ false beliefs? *Cognition*, 121 (3), 289-298. [10.1016/j.cognition.2011.07.011](https://doi.org/10.1016/j.cognition.2011.07.011)
- Luo, Y. & Baillargeon, R. (2005). Can a self-propelled box have a goal? *Psychological Science*, 16 (8), 601-608. [10.1111/j.1467-9280.2005.01582.x](https://doi.org/10.1111/j.1467-9280.2005.01582.x)
- Mayr, E. (1961). Cause and effect in biology. *Science*, 134 (3489), 1501-1506. [10.1126/science.134.3489.1501](https://doi.org/10.1126/science.134.3489.1501)
- Millikan, R. G. (1984). *Language, thought and other biological categories*. Cambridge, MA: MIT Press.
- (1989a). Biosemantics. *The Journal of Philosophy*, 86 (6), 281-297.
- (1989b). In defense of proper functions. *Philosophy of Science*, 56 (2), 288-302.
- (1990). Seismograph reading for “Explaining Behavior”. *Philosophy and Phenomenological Research*, 50 (4), 807-812.
- (1995). Pushmi-Pullyu Representations. *Philosophical Perspectives*, 9, 185-200.
- (1998). Conventions made simple. *The Journal of Philosophy*, 95 (4), 161-180.
- (2000). *On clear and confused ideas: An essay about substance concepts*. Cambridge, UK: Cambridge University Press.
- (2004). *Varieties of meaning*. Cambridge, MA: MIT Press.
- (2005). *Language: A biological model*. Oxford, UK: Oxford University Press.
- (2006). Reply to Tomasello. In M. Mazzone (Ed.) *Symposium on language: A biological model by Ruth Millikan*,. SWIF Philosophy of Mind Review.
- (2011). Reply to Neander. In D. Ryder & K. Williford (Eds.) *Millikan and her Critics* (pp. 37-41). Oxford, UK: Blackwell.
- Neander, K. (1991). Functions as selected effects. *Philosophy of Science*, 58 (2), 168-184.
- (1995). Misrepresenting & malfunctioning. *Philosophical Studies*, 79 (2), 109-141. [10.1007/BF00989706](https://doi.org/10.1007/BF00989706)
- (2004). Teleological theories of mental content. *Stanford Encyclopedia of Philosophy*. <http://plato.stanford.edu/entries/content-teleological/>
- (2007). Biological approaches to mental representation. In M. Matthen & C. Stephens (Eds.) *Handbook of the philosophy of science: Philosophy of biology* (pp. 548-565). Elsevier.
- (2011). Towards an informational semantics. In D. Ryder, J. Kinsbury & K. Williford (Eds.) *Millikan and her critics* (pp. 21-36). Oxford, UK: Blackwell.
- Onishi, K. & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, 308 (5719), 255-258. [10.1126/science.1107621](https://doi.org/10.1126/science.1107621)
- Origi, G. & Sperber, D. (2000). Evolution, communication and the proper function of language. In P. Carruthers & A. Chamberlain (Eds.) *Evolution and the human mind: Language, modularity and social cognition* (pp. 140-169). Cambridge, UK: Cambridge University Press.
- Perner, J. & Ruffman, T. (2005). Infants’ insight into the mind: How deep? *Science*, 308 (5719), 214-216. [10.1126/science.1111656](https://doi.org/10.1126/science.1111656)
- Perry, J. (1986). Thought without representation. *Proceedings of the Aristotelian Society*, 60 (137), 137-152.
- Piaget, J. (1954). *The construction of reality in the child*. New York, NY: Basic Books.
- Pietroski, P. (1992). Intentionality and teleological error. *Pacific Philosophical Quarterly*, 73 (3), 267-282.
- Recanati, F. (2002). Does communication rest on inference? *Mind and Language*, 17 (2), 105-126. [10.1111/1468-0017.00191](https://doi.org/10.1111/1468-0017.00191)
- (2007). Millikan’s theory of signs. *Philosophy and Phenomenological Research*, 75 (3), 674-681. [10.1111/j.1933-1592.2007.00103.x](https://doi.org/10.1111/j.1933-1592.2007.00103.x)
- Scott, R. M., Baillargeon, R., Song, H. & Leslie, A. (2010). Attributing false beliefs about nonobvious properties at 18 months. *Cognitive Psychology*, 61, 366-395.

- Scott-Philipps, T. C., Dickins, T. E. & West, S. A. (2011). Evolutionary theory and the ultimate-proximate distinction in the human behavioral sciences. *Perspectives on Psychological Science*, 6 (1), 38-47. [10.1177/1745691610393528](https://doi.org/10.1177/1745691610393528)
- Senju, A., Southgate, V., Snape, C., Leonard, M. & Csibra, G. (2011). Do 18-month-olds really attribute mental states to others? A critical test. *Psychological Science*, 22 (7), 878-880. [10.1177/0956797611411584](https://doi.org/10.1177/0956797611411584)
- Shea, N. (2007). Consumers need information: supplementing teleosemantics with an input condition. *Philosophy and Phenomenological Research*, 75 (3), 404-435. [10.1111/j.1933-1592.2007.00082.x](https://doi.org/10.1111/j.1933-1592.2007.00082.x)
- Sober, E. (1984). *The Nature of selection: Evolutionary theory in philosophical focus*. Chicago, IL: Chicago University Press.
- Southgate, V., Senju, A. & Csibra, G. (2007). Action anticipation through attribution of false belief by two-year-olds. *Psychological Science*, 18 (7), 587-592. [10.1111/j.1467-9280.2007.01944.x](https://doi.org/10.1111/j.1467-9280.2007.01944.x)
- Southgate, V., Chevallier, C. & Csibra, G. (2009). Sensitivity to communicative relevance tells young children what to imitate. *Developmental Science*, 12 (6), 1013-1019. [10.1111/j.1467-7687.2009.00861.x](https://doi.org/10.1111/j.1467-7687.2009.00861.x)
- Sperber, D. (2000). Metarepresentations in an evolutionary perspective. In D. Sperber (Ed.) *Metarepresentations: A multidisciplinary perspective* (pp. 117-137). Oxford, UK: Oxford University Press.
- (2001). An evolutionary perspective on testimony and argumentation. *Philosophical Topics*, 29 (1/2), 401-413. [10.5840/philtopics2001291/215](https://doi.org/10.5840/philtopics2001291/215)
- Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G. & Wilson, D. (2010). Epistemic vigilance. *Mind and Language*, 25 (4), 359-393. [10.1111/j.1468-0017.2010.01394.x](https://doi.org/10.1111/j.1468-0017.2010.01394.x)
- Sperber, D. & Wilson, D. (1986). *Relevance, communication and cognition*. Cambridge, MA: Harvard University Press.
- Tomasello, M. (2006). Conventions are shared. In M. Mazzone (Ed.) *Symposium on language: A biological model by Ruth Millikan*. SWIF Philosophy of Mind Review. <http://lgxserver.uniba.it/lei/mind/swifpmr.htm>
- (2008). *Origins of human communication*. Cambridge, MA: MIT Press.
- Tomasello, M., Call, J. & Hare, B. (2003). Chimpanzees understand psychological states – the question is which ones and to what extent. *Trends in Cognitive Sciences*, 7 (4), 153-156. [10.1016/S1364-6613\(03\)00035-4](https://doi.org/10.1016/S1364-6613(03)00035-4)
- Tomasello, M., Carpenter, M., Call, J., Behne, T. & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28, 675-691.
- Topal, J., Gergely, G., Miklosi, A., Erdohegyi, A. & Csibra, G. (2008). Infant perseverative errors are induced by pragmatic misinterpretation. *Science*, 321 (1269), 1831-1834. [10.1126/science.1176960](https://doi.org/10.1126/science.1176960)
- Warneken, F. & Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *Science*, 311 (5765), 1301-1303. [10.1126/science.112144](https://doi.org/10.1126/science.112144)
- (2007). Helping and cooperation at 14 months of age. *Infancy*, 11 (3), 271-294. [10.1111/j.1532-7078.2007.tb00227.x](https://doi.org/10.1111/j.1532-7078.2007.tb00227.x)
- Wellman, H. M., Cross, D. & Watson, J. (2001). Meta-analysis of theory of mind development: The truth about false belief. *Child Development*, 72 (3), 655-684. [10.1111/1467-8624.00304](https://doi.org/10.1111/1467-8624.00304)
- Wilson, D. & Sperber, D. (2004). Relevance theory. In L. Horn & G. Ward (Eds.) *The handbook of pragmatics*. Oxford, UK: Blackwell.
- Wimmer, H. & Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13 (1), 103-128. [10.1016/0010-0277\(83\)90004-5](https://doi.org/10.1016/0010-0277(83)90004-5)
- Woodward, A. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, 69 (1), 1-34. [10.1016/s0010-0277\(98\)00058-4](https://doi.org/10.1016/s0010-0277(98)00058-4)
- Wright, L. (1973). Functions. *The Philosophical Review*, 82 (2), 139-168.

Communicative Agency and *ad hominem* Arguments in Social Epistemology

A Commentary on Pierre Jacob

Marius F. Jung

A central point in Jacob's paper focuses on the incompatibility of Grice and Millikan's account of communicative agency. First, the Gricean mindreading thesis is incompatible with Millikan's direct perception account. Second, the account of cooperative devices, defended by Millikan, contradicts the Gricean separability thesis in a broad sense. While I agree with Jacob that these positions are indeed incompatible, I will shift focus and concentrate on issues concerning social epistemology with regard to communicative agency. A main issue in social epistemology concerns the accessibility of the speaker's reliability. How could the hearer remain epistemically vigilant without using fallacious reasoning? (i) I argue that the hearer, in order to be epistemically vigilant, could commit a local *ad hominem* attack, a process of inductive Bayesian reasoning which is an epistemic tool for assessing the speaker's reliability. (ii) Compared to this, a global *ad hominem* attack is a fallacious kind of reasoning, because it undermines knowledge transmission and it cannot be calculated in *Bayes' Theorem*. (iii) The account of a local *ad hominem* attack fits with Grice's mindreading thesis, which is incompatible with Millikan's account of direct perception. (iv) The Gricean separability thesis could better explain occurrences of *ad hominem* attacks than Millikan's assumption that speaker and hearer are cooperative devices.

Keywords

Bayesian reasoning | Communicative intention | Cooperative devices | Direct perception | Epistemic injustice | Epistemic vigilance | Global *ad hominem* argument | Informative intention | Local *ad hominem* argument | Mindreading | Personal attack | Positive reasons | Separability thesis

1 Introduction: Grice's individualistic account of meaning and epistemic trustworthiness

One of the main findings of Jacob's paper is a detailed elaboration of the differences between Millikan's (1984, 2004, 2005) communicative agency and the Gricean (Grice 1957, 1969, Sperber & Wilson 1986) account of speaker's meaning and intention. Jacob argues that the Gricean mindreading thesis, the separability

thesis, and the ostensive nature of communication are not supported by Millikan's account of the direct perception of speaker's intention, which supports a non-inferential model of the understanding of intentional signs. Furthermore, the Gricean account is incompatible with Millikan's claim that speaker and hearer are co-

Commentator

Marius F. Jung

mjung02@students.uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Target Author

Pierre Jacob

jacob @ ehess.fr
Institute Jean Nicod
Paris, France

Editors

Thomas Metzinger

metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu
Monash University
Melbourne, Australia

operative devices; the claim that the prediction of another's behavior could be explained through reliance on socially established conformities and conventions and that modern developmental psychology could get along without any theory of mind.

A very influential account of naturalizing the content of intentional mental representations is Millikan's teleosemantic framework (Jacob 2010). According to this view, the content of intentional mental representations can best be naturalized by relying on the history of the biologically-selected functions of these representations, namely the *direct proper functions* (cf. Millikan 1984, 1989). Interestingly, Jacob focuses on Millikan's concept of communicative agency, which is strongly connected to the teleosemantic framework, and argues that there are several aspects of the Gricean individualistic account of meaning that are more plausible than Millikan's when it comes to explaining social communicative agency. The most illuminating finding of Jacob's paper is the modern and precise presentation of the actuality of the Gricean separability thesis and the mindreading account, because it is explanatorily fruitful not only for philosophy of language, but also for social cognition, social epistemology, informal logics, and the relation between these different studies.

I generally agree with Jacob's main findings, nevertheless I will address some further issues of Millikan and Grice's account with regard to philosophical problems in social epistemology. Before I respond to these in detail, I first focus on the Gricean account and its implications for social epistemology.

The well-known Gricean (Grice 1957) account of the meaning of an utterance focuses on analysis of the speaker's meaning in a conversation. A speaker *S* means something *unnatural* if she intends something by the utterance of a sentence.¹ Let us suppose that the speaker is a politician with a specific agenda and with a propensity for aggressive propaganda. She utters the following:

(1) Speaker: "Our party will ensure that taxes go down".

The speaker *S* means (1) iff *S* utters (1) with the intention that a hearer *H* will gain the belief that *S*' party will make sure that taxes go down, if (i) the hearer *H* recognizes the speaker's intention (1) and (ii) because of that she gains the belief that *S*' party will make sure that taxes go down, (iii) since she recognizes that the speaker's intention is exactly that (cf. Grice 1957, 1969).

Since Gricean meaning is individualistic and subjective it is important to note which underlying cognitive states constitute this meaning. As Jacob puts it—relying on Sperber & Wilson's (1986) interpretation of the Gricean account²—there are three main assumptions upon which the psychological theory of meaning is based, namely the separability thesis, the mindreading thesis, and the asymmetry between an informative and communicative intention. Together with Jacob I shall focus on Sperber and Wilson's account, which argues that the Gricean theory can be summarized as a reciprocal process of intentions. The *informative intention* is an intention of a speaker who wants to inform a hearer about some state of affairs. In order to be successful, the speaker has also the intention that the hearer recognize the informative intention (Grice 1957). This means that at first the hearer has to understand the informative intention. If she understands it, the *communicative intention* of the speaker has been fulfilled. But the informative intention will only be fulfilled, if the speaker is trustworthy: a necessary condition for accepting a speaker's utterance. In effect, the hearer gains a new belief. The assessment of her trustworthiness depends on the hearer of that intention (Sperber & Wilson 1986). She must admit that the speaker has to be reliable in order to be trustworthy, or, to put it in Jacob's (this collection, p. 4) words, "the addressee must further accept the speaker's epistemic or practical authority". But a question arises: on which kind of epistemic practices does the hearer have to rely in order to accept the

1 Grice (1957) distinguishes between a *natural* and an *unnatural meaning*. Unnatural meaning is always characterized by the speaker's intention. The natural meaning of a sign characterizes meaning that is independent of a speaker.

2 I follow Jacob in relying partly on Sperber & Wilson (1986) when I talk about the Gricean account.

speaker as an epistemic authority? I will address this question in the following commentary.

In order to answer it, I will argue that (i) a hearer could commit a *local ad hominem attack*, a process of inductive *Bayesian reasoning* that secures epistemic vigilance. Roughly, an *ad hominem* attack is an argument that considers rather personal properties of an utterer than the argument itself. (ii) A fallacious kind of the personal attack is the *global ad hominem attack*, which undermines every testimony of a speaker because of its personal traits. (iii) The Gricean account of mindreading could better account for an inductive inference model than Millikan's direct-perception-account. (iv) Practices of *ad hominem* attacks, I will argue, support the Gricean separability thesis, while Millikan's co-operative devices account is less plausible.

The structure of this commentary will be as follows: first, I focus on Grice and Millikan's framework and its implications for social epistemology, namely the problem of epistemic reliability (cf. section 2). In section 3 I shall present Lackey's account of a social epistemological dualism, a hybrid theory in which Lackey tries to connect the most plausible findings of social epistemological reductionism and anti-reductionism. Then I argue that the Gricean account of informative intentions and Lackey's *positive reason component* could lead to the *personal attack* or *ad hominem argument* (cf. section 4). In section 5 I argue that there are two possible commitments of *ad hominem* arguments, to be specific, the *global* and the *local ad hominem attack* (cf. section 5.1, section 5.2). In the Gricean account of the *mindreading thesis* is compatible with the drawn picture of our social epistemological practices, because it supports the inductive inference model, while Millikan's account of *direct perception* could not account for this. The Gricean *separability thesis* fits nicely with the positive reasons component and the reliance on *ad hominem* arguments, while

Millikan's account of speaker and hearer as *co-operative devices* is less plausible (cf. section 6).

2 Epistemic intentions and epistemic reliability

The utterance of a speaker depends on two directions of fit. The first can be characterized as a *mind-to-world-relation*. Here, the speaker has the intention of conveying some states of affairs about the actual world. This direction of fit implies that the speaker wants to share some epistemic notions. If she is successful in doing so, the hearer will gain a true belief. This class of utterances is descriptive.

The second is a *world-to-mind direction* of fit of the speaker's utterance. Here the speaker wants to convey some of her desires to the hearer, who acts in a particular way in order to fulfill the speaker's desire. The intention is fulfilled if the hearer gains a new desire to act in order to fulfill the speaker's desire (Sperber & Wilson 1986). This kind of direction of fit is unimportant for the following account. Here I shall focus on descriptive utterances.

Before I address some implications of epistemic intentions, I shall focus on the *separability thesis*. This thesis addresses the problem of an asymmetry of interests between hearer and speaker. Since the interests are not identical, the speaker could deceive the hearer. And the other way around: the hearer could distrust the speaker even though she utters a true sentence. Sperber et al. (2010) claim that some amount of distrust is a stabilizer in the evolution of human communication, which they call *epistemic vigilance*. Imagine that humans believed almost everything they were told. Since not every speaker has the propensity to speak the truth, hearers would have a lower amount of knowledge because they would have no tool for distinguishing a reliable testimony from a non-reliable one. Communication would be very imprecise, because knowledge agency would be less successful. Hence, epistemic vigilance is a precondition for cooperative communication, because both speaker and hearer check the reliability of knowledge transition. I agree with Sperber et al. (2010) that epistemic vigilance is a

feature of a *source* and the *content* of information.³

Let me go back to the mind-to-world-relation of fit. If the twofold account given by Sperber & Wilson (1986) is correct, the hearer has gained a true belief (about some state of affairs). To count as knowledge, we have to ask whether the true belief is justified.⁴ What could count as a justification? Some social epistemologists would say that the testimony of the speaker is sufficient to count as a justification. This is the thesis of an *anti-reductionism* in social epistemology which contains the claim that the speaker must not rely on other sources of knowledge such as perception, inference or memory to justify her belief (Coady 1992). A *reductionist* would say that the testimony cannot count as knowledge without relying in addition upon other sources of knowledge (Fricker 1995).

In Millikan's (2005) account of an intentional conventional sign (which is the content of an intentional mental representation), she assumes that speaker and hearer are cooperative devices that have co-evolved. The relationship between sender and receiver can be characterized as beneficial in the long term. Millikan proposes a framework, in which the descriptive representations describe a mind-world-direction, whereas the *directive* representation describes the world-to mind relation. The long term beneficial communicative agency between sender and receiver characterizes a function of reproduction of conventional signs. The *direct proper function* in this particular case is that the hearer gains a new belief. Millikan (1984) is well known for this teleosemantic account that can deal with misrepresentations. In such a case, the proper function remains unfulfilled. It is unfulfilled if the speaker fails to cause a new belief in the hearer (Millikan 1984, 2005). But the hearer could also be responsible for the unfulfilled proper function if she mistakenly judges the speaker to be untrustworthy. The hearer is also

a constitutive part of the cooperative devices that establish the direct proper function (Millikan 1984, 2005).

3 Social epistemology: Lackey's dualism

Before I present a more detailed account of *ad hominem arguments*, I will say a few words about social epistemology and the position that is presupposed in this commentary. Jennifer Lackey's (2006) account of social epistemology relies upon a kind of dualism, in which she combines anti-reductionism and reductionism. According to her, social epistemology has made the mistake of addressing the debate between reductionism and non-reductionism unilaterally. *Reductionism* takes epistemic responsibility and the rationality of the hearer far too seriously, because the hearer has to rely upon other sources of knowledge like perception, memory, deductive inferences, etc. The claim here is that testimony is not a source of knowledge in the first place, because a hearer could never know the intentions of a speaker who held accidentally or intentionally false beliefs. In contrast, *anti-reductionism* always focuses on the speaker's perspective and her propensity for credible testimonials. Proponents of anti-reductionism claim that a large amount of our knowledge depends on testimonials. We would know almost nothing if we were as restrictive as the reductionist claims (Coady 1992). Lackey wants to combine these two accounts in a kind of dualism. Her dualism contains the presupposition of the reliability of the speaker along with *positive reasons* to accept the speaker's testimony, evaluated from the hearer's perspective. If the speaker utters a true sentence and the hearer has positive reasons to trust the speaker, then knowledge from testimony is possible. Lackey argues for the following conditional, which contains three necessary conditions:⁵

For every speaker A and hearer B, B justifiably believes that p on the basis of A's testimony that p only if: (1) B believes that p on the basis of the content of A's

³ I will address this topic with respect to the *ad hominem* argument in section 5.

⁴ I will not address Gettier cases with regard to social epistemology. For the sake of this commentary, I will use the term *knowledge* as meaning justified true belief. Issues concerning the *ad hominem* fallacy will concern the justification-condition.

⁵ Lackey claims that dualism accounts only for necessary conditions for a source of knowledge.

testimony that *p*, (2) A's testimony that *p* is reliable or otherwise truth conducive, and (3) B has appropriate positive reasons for accepting A's testimony that *p*. (2006, p. 170)

For the present account it is important that a testimony, given by A, qualifies as a source of knowledge that depends on the hearer having positive reasons to think that A's testimony is reliable. Recall the informative intention and the mind-to-world-direction of fit, which is fulfilled if the speaker causes a new belief in the hearer. The direct proper function of the co-operation between speaker and hearer in Milikan's (1984, 2005) account would be fulfilled. But according to Lackey's condition (3), the achievement of a new belief is only justified if there are various *positive reasons* that account for the reliability of the speaker's testimony. Consider the account of Sperber et al. (2010, p. 379) that "the filtering role that epistemic vigilance [...] in the flow of information in face-to-face interaction" is an important feature of communicative agency. But which kind of filtering do they mean? In other words, what are positive reasons, exactly? Could they be past experiences about the reliability of the speaker or even a group to which the speaker belongs?

4 *Ad hominem* arguments and epistemic injustice

I claim that the Gricean account of communication supports our social practices of committing *ad hominem* arguments. The committing of *ad hominem* attacks in communicative agency becomes patent when you look at *positive reasons* in more detail. During past events of communicative agency, a hearer has tested the trustworthiness of several speakers on the basis of her personal properties and the context to which these properties have been related (Lackey 2006; Fricker 2007). The ability of being epistemically vigilant emerges very early in human development. At the age of three years, infants already prefer testimony from a reliable source (cf. Clément 2010). From that age on, infants develop a "cognitive filter that enables children

to take advantage of testimony without the risk being completely misled" (Clément 2010, p. 545).⁶

Now back to the example: suppose that the hearer in question has been confronted with the testimony of politicians in the past. She then hears the following sentence from speaker *S*:

(1) "Our party will make sure that taxes will go down".

Would you, as a hearer, believe her? Consider past cases of political propaganda and ask yourself how reliable the politician, the speaker, really is. At first, let us assume you do not. You are a very skeptical person, especially when it comes to political issues. Is it rational to be skeptical, so are you guilty of prejudice? Let us assume that the speaker is surprisingly reliable. She speaks the truth. The party wins and reduces taxes. Have you treated the politician in an epistemic inequitable way or was it the only way to remain epistemic vigilant? These questions will be addressed in the following sections.

The hearer who does not believe *S*' utterance (1) has committed an *ad hominem* argument, a personal attack against the speaker. According to Walton (2008, p. 170) "[t]he *argumentum ad hominem*, meaning 'argument directed to the man', is the kind of argument that criticizes another argument by criticizing the arguer rather than his argument." A hearer takes some personal properties, such as being a politician, and infers that the expressed sentence is false. It is *prima facie* irrelevant to consider personal traits as indicators of a false testimony *t* (Yarp 2013; Walton 1998).⁷ Keeping Walton in mind, we are able to generalize *ad hominem* attacks as follows:

⁶ There is further evidence in developmental psychology that speaks of very early acquisition and practice of epistemic vigilance (cf. Clément 2010; Sperber et al. 2010; Mascaro & Sperber 2009). Another issue with regard to the positive reasons component in social epistemology is the so-called *infant/child-objection*. This concerns the hearer's competence in evaluating the speaker's reliability that small children lack, which is often construed as an argument against reductionism (Lackey 2006). For a general discussion see Lackey (2005).

⁷ As will be seen in section 5.1, there are some exceptions where personal properties are relevant.

Ad hominem attack

- (1) Speaker *S* gives a testimony *t*.⁸
 - (2) The speaker's *S* property φ is a negative property with regard to trustworthiness.
 - (3) Speaker *S* has a negative property φ , which is ascribed as relevant for her testimony *t* by hearer *H*.
-
- (C) The testimony *t*, uttered by speaker *S*, is false as assessed from hearer *H*.⁹

The arguments of the speaker (implicitly represented or explicitly formulated) have not been challenged seriously by the hearer. She just considers personal traits sufficiently to reject the given argument or proposition.¹⁰ The allegedly suboptimal personal characteristics of the person do not provide any evidence for rejecting the proposition *p*. The hearer neither shows that the deduction of the speaker includes fallacious reasoning nor that the premises on which her proposition is based are wrong. Informal logic does not support the hearer in this situation (Groarke 2011). Has the speaker been treated inequitably? Miranda Fricker (2007) tries to answer this question and introduces the notion of *epistemic injustice*. She generalizes the notion as follows:

Any epistemic injustice wrongs someone in their capacity as a subject of knowledge, and thus in a capacity essential to human value; and the particular way in which testimonial injustice does this is that a hearer wrongs a speaker in his capacity as a giver of knowledge, as an informant. (Fricker 2007, p. 5)

⁸ I assume that a testimony *t* expresses an argument that contains the relevant proposition *p*.

⁹ One could of course distinguish between a testimony and an argument. Here I presuppose that a testimony is somehow a conclusion of an argument. Fricker (2007, p. 61) supports this view as follows: "One might be inclined to put a familiar picture of justification to the fore and argue that in order to gain knowledge that *p* from somebody telling her that *p*, the hearer must in some way (perhaps very swiftly, perhaps even unconsciously) rehearse an argument whose conclusion is *p*."

¹⁰ For the purpose of this commentary I will defend a weak view of propositions. The utterance of a speaker expresses a proposition that is true if it represents a state of affairs. There is of course an asymmetry between the propositional content of an utterance and the propositional content of the speaker's belief or thought. I agree with Jacob (1987) that it is sufficient to assume similarity between the two.

It fits Fricker's generalization that the capacity of a speaker to convey true beliefs is undermined. The positive reasons clause of Lackey's dualism also supports this step of reasoning because the character or the identity of a speaker could be relevant for her evaluation of trustworthiness in epistemic contexts (cf. Fricker 2007; Lackey 2006). Crucially, stereotypes and prejudices—based upon *ad hominem* arguments—are paradigmatic cases of epistemic injustices (cf. Fricker 2007). But is it not rational for a hearer to distrust our politician? Jacob (this collection, pp. 4–5) claims that "not every speaker is (or should be) granted equal epistemic or practical authority on any topic by every addressee." Remember that the hearer's positive reason component is a remainder of the reductionist account with regard to testimony as a justifier of knowledge. Is it not a necessary condition for the positive reason component to remain vigilant in such contexts? If epistemic vigilance does not play a role in this context, then Lackey's suggestion of the necessary condition of positive reasons on the hearer's side is implausible. In the following I shall argue that epistemic vigilance is very important and that the dualistic account could account for it. Nevertheless, one has to accept what I call a local *ad hominem* argument in order to be epistemically vigilant in our particular case.

5 Two kinds of *ad hominem* attack

In the following section I make a suggestion in order to disarm the problem of the *ad hominem* argument with regard to the *positive reason* component. Does an *ad hominem* attack always include fallacious reasoning? Walton (2008, p. 170) claims that "the *argumentum ad hominem* is not always fallacious, for in some instances questions of personal conduct, character, motives, etc., are legitimate and relevant to the issue." Even though cases of *ad hominem* arguments might sometimes be informally fallacious, there are some highly relevant cases in which a particular *ad hominem* attack could be committed in order to remain epistemic vigilant. Since you, as a hearer, have a set of positive reasons—for instance being aware of the usual verbal

espousals of politicians during an election campaign—you are forced to commit a personal attack. Did your reliability assessment rely on fallacious reasoning? In the literature there is a common distinction between three types of the *ad hominem* argument: the *abusive*, the *circumstantial*, and the *tu quoque* argument (Groarke 2011; Walton 1998, 2008).¹¹ These kinds of personal attacks describe various pragma-dialectical reasoning in interpersonal communicative relationships. Below (cf. section 5.1, 5.2) I want to draw a further distinction between two kinds of *ad hominem* attack that are closely connected to communicative situations explicitly involving knowledge transmission. Hence I want to provide a framework that fits well with social epistemological dualism.

Before this, I want to address the *Bayesian argumentation model*, which is presupposed by the following account of *ad hominem* arguments. Few things have been said about the inductive reasoning model which is the underlying mechanism of an *ad hominem* attack. Roughly, one has to consider past experiences with regard to reliability, constituted by contexts and speaker properties, to adjust this experience for future communication. Harris et al. (2012) provide an account that fits well with an inductive model of reasoning, because, potentially evidence is not provided by deductive reasoning. They claim that the evaluation of a proposition or a given testimony is based on an individual's probabilities, which could be described formally using *Bayes' Theorem*. A big advantage of this account is that it describes our subjective evaluations in daily experiences very well. Often, when we are asked “do you believe *S*?” we are inclined to say something like “I am not sure. I guess not”. This could be well explained with the Bayesian model, where an individual's belief does not have a truth-value of 0 or 1. The relevant belief is instead estimated in one's subjective degree of that belief, as a probability between 0 and 1. Let us embed this in our current considerations. The utterance type of the politician is already embedded in a

kind of bias or in posterior beliefs about politicians in general: this is called the hypothesis *h*, and has a particular probability $P(h)$ in isolation from evidence *e*. Evidence *e* in this particular case is constituted by personal characteristics, properties, and circumstances of the utterance. The receiving of the new evidence *e* should update $P(h)$, the probability of the hypothesis. Individuals ought, according to the normative stipulation, if they receive any evidence, to let it influence the probability of the proposition: “[this] normative procedure by which individuals should update their degree of belief in a hypothesis *h* upon receipt of an item of evidence *e* is given by Bayes' Theorem:

$$P(h|e) = \frac{P(h)P(e|h)}{P(e)} \quad \text{“(Harris et al. 2012, p. 316)”}$$

$P(h|e)$ describes the conditional probability of a hypothesis being true after one has received evidence *e*. $P(e|h)$ terms the conditional probability of receiving evidence *e*, given hypothesis *h*. $P(e)$ just describes the evidence in isolation from the truth-values of the hypothesis *h* (cf. Harris et al. 2012).

5.1 Local *ad hominem* attack

I am now able to distinguish two kinds of *ad hominem* attack, a local and a global one. I will thus sketch out some considerations with regard to the presented Bayesian framework.

The positive reason component describes the practice of a speaker's credibility assessment. This process of credibility assessing could lead to what I will call a *local ad hominem* attack. Roughly, one commits a local *ad hominem* argument if one acknowledges someone to be trustworthy in general, but with some exceptions in particular cases. If you ask the politician what time it is or the straightest way to the subway station, it is very unlikely that she would have the intention of deceiving you (Sperber 2001). Hence, one would count her as a trustworthy person. Nonetheless, given the particular information about her party and the reduction of taxes, you might find it unlikely that she is telling the truth. As long as you do not dismiss her in general as an eligible bearer of

¹¹ For some very interesting empirical investigations with regard to these three kinds of *ad hominem* arguments, see van Eemeren et al. (2000) and van Eemeren et al. (2008).

knowledge, it is vigilant in a rational way to distrust her in this case. This view could be described as a subclass of epistemic vigilance, because it is an evolved tool that minimizes the risks of deception which is—according to Sperber et al. (2010)—a condition for cooperative communicators.

At this stage, subjective Bayesian probability comes into play. The general bias of your past experiences with regard to politicians, which enters the stage before you have received the evidence, could be described with (h) . The probability of the hypothesis $P(h)$ is the probability of your believing her without having received evidence e . Given (1), evidence e describes that the person is a politician during a campaign, which also has a particular probability, termed $P(e)$. The evidence condition is the part that divides the local *ad hominem* attack from the global, because the evidence is able to influence one's subjective degree of probability. The personal traits of the speaker as well as the context of utterance-use serve as evidence e . $P(h/e)$ is then the conditional probability of h , if the hearer H receives evidence e . Given the hypothesis h , the probability of receiving evidence e is described by $P(e/h)$. As presented above, this could be calculated within *Bayes' Theorem* (cf. Harris et al. 2012).

As you can see, the *content* and the *source* of information, which serve as evidence e , are both similarly relevant for a testimony (cf. Sperber et al. 2010). Buening (2005) calls this kind of reasoning *relevance-based* with regard to the relevant circumstances that could invoke *ad hominem* attacks. The context and content of the utterance or proposition in question is important for assessment. Walton (1998) calls such a context and content-related *ad hominem* attack a *credibility function*. The proposition in question undergoes an *ethotic rating*, a kind of evaluation of a person's epistemic input value (of her testimony), which can go up or down. When committing a local *ad hominem* attack, the rating goes down. Hence the credibility of the speaker is undermined in a specific case that affects the proposition, which fits with the account of Bayesian argumentation. The local *ad hominem* argument is an example of a non-

fallacious *ad hominem* attack because it is *content* as well as *context-related* and epistemically equitable. The normative stipulation of the Bayesian account that a hearer “should update [her] probabilistic degrees of belief in a hypothesis in accordance with the prescriptions of Bayes' Theorem” (Harris et al. 2012, p. 316) is fulfilled in the local version. The take home message of this passage could be presented in a more simplified way:

Local *ad hominem* attack (non-fallacious)

- (1) Speaker S gives a testimony t .
 - (2) Speaker's S property ϕ is a negative property with regard to trustworthiness.
 - (3) Speaker S has a negative property ϕ that is relevant evidence e for a hypothesis h with regard to the content of the particular testimony t by hearer H .
-
- (C) The testimony t , uttered by speaker S , is probably false as assessed by hearer h .¹²

5.2 Global *ad hominem* attack

I call the opposite kind of reasoning the *global ad hominem argument*, which I claim, is fallacious. The hearer commits a global attack if she does not believe the speaker in general. The hearer discredits her any kind of trustworthiness. Consider some stereotypes and prejudices that could suffice for such a radical conclusion.¹³ This behavior is clearly irrational, since it undermines any testimony of a speaker in every situation. As you can see, this kind of fallacy is neither a *content-related* nor *context-related ad hominem* argument. In distinction from the global attack, here the content does not play any role in the evaluation of the speaker's reliability. First, the evidence e only includes per-

¹² As Jacob (1987) suggests, beliefs are shared in different communities with different ideological backgrounds that are themselves constitutive of belief-formation. One could defend that the local *ad hominem* attack is an important tool for running a communicative society. If so, we would be using local *ad hominem* attacks as a form of self-deception, which would then be somehow an instance of a shared optimism bias. I will not discuss this phenomenon any further, because it is not a tool or cognitive filter that improves knowledge transmission. Hence, positive local *ad hominem* attacks, one could argue, have at least a propensity for being epistemically unvigilant mechanisms.

¹³ Yarp (2013) suggests that *ad hominem* fallacies like prejudices could be unconscious or at least not transparent to the hearer's reasoning.

sonal traits and not the circumstances of the utterance. Second, the evidence e does not influence the degree of belief that the hypothesis h is true. This is the reason why the belief, formed via the process of a global *ad hominem* fallacy, could not be calculated in *Bayes' Theorem*. The normative stipulation that evidence e should affect the probability of a hypothesis h is not satisfied.

Another reason why this type of personal attack is fallacious is because it includes irrelevant circumstances and personal traits as the basis of the speaker's evaluation (Buenting 2005). The speaker, as she is assessed, is not in any way disposed to maximizing the hearer's set of true beliefs.¹⁴ In other words, the hearer undermines any potential benefit she may gain through any of the speaker's testimony (Sperber 2001), which could be described as a paradigmatic case of epistemic injustice (Fricker 2007). Hence, a general assessment of the speaker's credibility has nothing to do with the process of positive reason formation. In other words, the global *ad hominem* attack is not an epistemic tool, because it lacks any *credibility function*, because the speaker assesses the testimony as necessarily false, because of her personal properties. Since it is not an epistemic tool, this kind of reasoning is not epistemically vigilant (Sperber et al. 2010). To summarize:

Global *ad hominem* attack (fallacious)

- (1) Speaker S gives a testimony t .
- (2) Speaker's S property φ is a negative property with regard to trustworthiness.
- (3) Speaker S has a negative property φ that is ascribed as relevant for every testimony t by hearer H .

(C) The testimony t , uttered by speaker S , is necessarily false as assessed by hearer H .

6 *Ad hominem* arguments and communicative agency

I agree with Jacob that some aspects of Grice's theory of meaning are in a broad sense incompatible with Millikan's account of communicat-

ive agency. The focus of this commentary so far has been communicative agency in epistemic contexts and its implications, and in particular the personal attack. I will now evaluate whether Millikan's account of direct perception or Grice's account of mindreading could account for *ad hominem* arguments in epistemic contexts. My answer is that the Gricean mindreading thesis is more plausible. I then compare the separability thesis with the cooperative devices. The separability thesis fits best with the practices of *ad hominem* fallacies. The presupposition of cooperative devices is less plausible.

6.1 Mindreading vs. direct perception

Recall from section 1 that the mindreading thesis relies on the twofold account of informative and communicative intention. First, the speaker has to recognize or understand the speaker's informative intention, which is the speaker's communicative intention. For the *fulfillment* of the informative intention, the trustworthiness of the speaker has to be accepted. In other words, in order to *fulfill* the informative intention, the hearer commits neither a local nor a global *ad hominem* argument (or she would not accept it). But in order to be an epistemically vigilant agent, the hearer has to make some further inferences, which are inductive (as well as the *ad hominem* fallacies). This inductive inference model involves some kind of mindreading that could affect the reliability judgment.¹⁵ Millikan claims that the acceptance of a given testimony as a source of knowledge is a form of direct perception without any kind of inference (Millikan 1984; Sperber et al. 2010). She talks about *translation* instead of inference. The hearer translates the utterance via direct perception into a new belief (Millikan 2004):

Forming a belief about where Johnny is on the basis of being told where he is I just as direct a process (and just as indirect) as forming a belief about where Johnny is on

¹⁴ I will not consider the ethical implications of this view any further in this commentary.

¹⁵ Unfortunately, I cannot address in this paper which kind of mindreading is supported by this view and how it could perhaps be related to social cognition and mirror-neurons. For a general discussion see Jacob (2008, 2013).

the basis of seeing him there. (Millikan 2004, p. 120) There is no reason to suppose that any of these ways of gaining the information that Johnny has come in requires that one perform inferences. (Millikan 2004, p. 125)

It is doubtful that these circumstances explain our everyday communicative agency, especially with regard to epistemic conversations. According to Millikan, the acceptance of a new belief does not involve any representation of the speaker's intention. But in order to assess the reader as benevolent and competent (or reliable), one has to rely—as argued in section 5—on inductive inferences which are of course derived representations manifested in beliefs about the speaker's intention.

In epistemic contexts of communication only the mind-to-world direction is involved, *qua* descriptive utterances. One criticism offered by Jacob is that Millikan's account of perception could only account for descriptive utterances, hence only for the mind-to-world direction. Another issue is closely related to this kind of criticism. It concerns testimony that has very little to do with perceptual capacities. With regard to very complex utterances like (1), I agree with Jacob (this collection, p. 9) that “it does not make much sense to assume that either the speaker or her addressee could perceive what the speaker's utterance is about.” Consider the nature of testimonial reports. Even some direct perception of a testimony about some state of affairs is perceptually impoverished compared to directly perceiving the state of affairs in question. Imagine some testimonial reports that have been heard through the radio. In such a case, you are not in a perceptually close relationship to the reported state of affairs. If you evaluate the credibility of the speaker, it is very likely that you would run through different processes of inductive inference in order to commit an *ad hominem* argument or avoid one. The more abstract the testimony, the more implausible it becomes that it has anything to do with direct perception. It becomes even more complicated with complex indexical utterances or a group of different but

equally eligible interpretations of a particular testimony. Consider again example (1). Here it is very likely that a hearer represents some intentions of the speaker that are linked to her psychological states. If one representation is that the speaker could deceive the hearer in particular circumstances, the hearer will probably commit a local *ad hominem* attack. To sum up: *Ad hominem* arguments are ascriptions that result from inductive inferences that also depend on belief-desire psychology, because the hearer gains a representation of the second-order representation of the sentence expressed by the speaker. The representation of the hearer is a third-order representation of the second-order linguistic representation of the speaker (cf. Jacob 1987).¹⁶

6.2 Separability thesis vs. cooperative devices

The problems addressed so far are closely related to the separability thesis. The separability thesis is the claim that the hearer and the speaker could have different interests, which are causes of the informative intention remaining unfulfilled, because there are two cases that suggest that the interests of both parties fall apart. In the first, the hearer gains a new belief that is not true, because the speaker has the informative intention to deceive the hearer. So her informative intention has the aim that the speaker gains a false belief and not one about some states of affairs, as described in section 2. In the second, the sentence, uttered by the speaker, is true, but nonetheless denied by the hearer on the basis of an *ad hominem* argument (Sperber & Wilson 1986).¹⁷ These two cases do not support Millikan's (2005) claim that the interests of both speaker and hearer are balanced. If a hearer commits a global *ad hominem* argument, it is even harder to ascribe balanced interests to speaker and hearer. Sperber (2001) defends a plausible weak version of coincidence of interests. It is only necessary that they over-

¹⁶ According to Jacob (1987), a belief-ascription is not constitutive of the subject's belief in the first place.

¹⁷ There are, of course, plenty of other options for different interests (cf. Sperber 2001).

lap in the long term in order to establish successful practices of social knowledge transmission.

Cases of global *ad hominem* arguments could only occur if a speaker understands the informative intention which she combines with some particular personal properties of the speaker in order to reject the testimony in question. Hence the speaker succeeds in establishing the communicative intention, but fails to fulfill the informative intention. In Millikanean terms, the direct proper function of the speaker is that the hearer gains a new belief. If the hearer commits an *ad hominem* attack, the direct proper function remains unfulfilled. But the communicative intention is still fulfilled, and that is all that is required for successful communication according to the separability thesis of communicative intentions (Sperber & Wilson 1986). The hearer recognizes that the speaker wants to inform her of her informative intention, which means that the communicative intention has been fulfilled. But the informative intention—which is that the hearer gains some new information or a true belief—fails, because an *ad hominem* attack has been committed. This circumstance could be well explained with the separability thesis and the weak account of communication that we addressed in section 6.1. To sum up, and in agreement with Jacob, if an *ad hominem* attack has been committed, even a weaker version, communicative agency is violated in the Millikan (2004, 2005) framework because the cooperative conventional transmission is violated in the first place.¹⁸

The picture I draw with regard to the *ad hominem* arguments rests on the assumption that trustworthiness has to be assessed by the hearer in order to be counted as epistemic vigilant, which would be the reductionist component. On the other side, the speaker has to utter a true sentence to transmit knowledge to a speaker, which would be the anti-reductionist-component. This view is supported by Lackey's dualism. As can be seen, the establishment of a dualistic account and all its implications for the

inductive reasoning model can be better explained with the separability thesis.

7 Conclusion

In this commentary I have extended the refreshing account given by Jacob, who presents Grice's individual account of meaning and assesses its plausibility with regard to communication and knowledge transmission. I defended the view that one promising way to talk about testimony as a source of knowledge is offered by Lackey's dualism. Here, both speaker and hearer are the important in knowledge transmission. In order to secure this transmission, the speaker has to utter a true sentence and the hearer has to check the speaker's trustworthiness. I distinguished two kinds of personal attack:

(i) The local *ad hominem* argument, which is not fallacious, focuses on the proposition and personal properties of the speaker, and is a content-related, relevance-based attack based upon one's subjective probabilistic estimation of the speaker's reliability, which can be calculated in *Bayes' Theorem*.

(ii) The global *ad hominem* argument, which is fallacious, is an extreme prejudice that denies that the speaker is reliable in any case. It is not usable as a tool for knowledge transmission, because it violates the stipulation of the Bayesian argumentation in which one should include some evidence in the subjective probability estimation. This extreme kind of a personal attack could be racism or stigmatizing, for instance.

(iii) Here I argued—in agreement with Jacob—that the Gricean account of mindreading is more plausible than Millikan's account of direct perception. To use the inductive model of reasoning when evaluating a speaker's reliability, one also has to rely on the use of belief-desire psychology. It is important to *infer* the intentions of the speaker, and to think about whether she has good reasons or a general propensity to speak the truth. Millikan's framework of direct perception does not account for this, because the direct perceptibility of some

¹⁸ The question, of course, is in which sense it is violated, in detail, and how this affects Millikan's theory of language in general. However, these implications cannot be addressed here.

abstract cases of testimony and their evaluated reliability is very implausible.

(iv) Last, I argued that the description of speaker and hearer as cooperative devices is implausible, too. First, we know that the speaker could have deceptive intentions. Second, if somebody is committing a global *ad hominem* attack, the interests of the utterer and her addressee fall apart. Nonetheless, the communicative intention still holds. Both parties communicate successfully, even if the hearer does not gain a new belief. So we could conclude with Grice and Sperber and Wilson that the communicative intention is sufficient for a successful communication. If the addressee commits an *ad hominem* fallacy, the proper function is unfulfilled. But in this case the conventional speaker action, which is part of successful communication, has been violated. I have argued that the Gricean account could well explain our communicative practices regarding epistemic contexts.

In terms of future research, it would be very interesting to see, how the Gricean philosophy of a speaker's individual meaning and mindreading could be embedded in theories about social cognition, social epistemology and informal logics. Jacob has presented an illuminating account of how the Gricean philosophy could be embedded in modern philosophy of mind and the cognitive sciences. I propose that one should further reflect on Jacob's arguments and adopt his conceptual framework, which I think is very precise and explanatorily fruitful. But for my own proposal of a distinction between local and global *ad hominem* attacks, it will be important to flesh out these accounts with regard to Bayesian reasoning and argumentation in epistemic contexts. A good candidate for elaborating this kind of research is the recent account of *predictive processing*, which is also based on Bayesian probabilities.

In this commentary I have claimed that these personal attacks are inductive mechanisms. But much more could be said about their functionality or even their instantiation in a cognitive system. Then it would be interesting to see if non-human cognitive systems could commit these kinds of *ad hominem* attacks. How precise could they be in evaluating a

speaker's reliability? Are instances of *ad hominem* attacks bound to a specific type of brain through which the relevant representational and functional architectures are realized? How could such a phenomenon like the global and local *ad hominem* attack evolve in *homo sapiens*, and what are the deeper underlying cognitive mechanisms of such attacks? These questions need to be answered if we want to understand these important mechanisms and processes of social knowledge, as well as our communicative society as a whole.

Acknowledgements

First of all, I am appreciative for the illuminating target paper. In addition to that, I would like to thank the two anonymous reviewers and Thomas Metzinger and Jennifer M. Windt for their editorial reviews. They really helped to improve this paper. I am also grateful to Thomas Metzinger and Jennifer M. Windt for the opportunity to contribute to this project.

References

- Buenting, J. M. (2005). The rejection of testimony and the normative recommendation of non-fallacious 'ad hominem' arguments based on Hume's 'Of Miracles' and Canadian Law. *Auslegung*, 27 (2), 1-16.
- Clément, F. (2010). To trust or not to trust? Children's social epistemology. *Review of Philosophy and Psychology*, 1 (4), 531-549. [10.1007/s13164-010-0022-3](https://doi.org/10.1007/s13164-010-0022-3)
- Coady, C. A. J. (1992). *Testimony: A philosophical study*. Oxford, UK: Oxford University Press.
- Fricker, E. (1995). Telling and trusting: Reductionism and anti-reductionism in the epistemology of testimony. *Mind*, 104 (414), 393-411.
- (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford, UK: Oxford University Press.
- Grice, H. P. (1957). Meaning. *The Philosophical Review*, 66 (3), 377-388. [10.2307/2182440](https://doi.org/10.2307/2182440)
- (1969). Utterer's meaning and intentions. *The Philosophical Review*, 78 (2), 147-177. [10.2307/2184179](https://doi.org/10.2307/2184179)
- Groarke, L. (2011). Informal logic. *Stanford Encyclopedia of Philosophy*. <http://plato.stanford.edu/entries/logic-informal/>
- Harris, A. J. L., Hsu, A. & Madsen, J. K. (2012). Because Hitler did it! Quantitative tests of Bayesian argumentation using ad hominem. *Thinking & Reasoning*, 18 (3), 311-343. [10.1080/13546783.2012.670753](https://doi.org/10.1080/13546783.2012.670753)
- Jacob, P. (1987). Thoughts and belief ascriptions. *Mind & Language*, 2 (4), 301-325. [10.1111/j.1468-0017.1987.tb00124.x](https://doi.org/10.1111/j.1468-0017.1987.tb00124.x)
- (2008). What do mirror neurons contribute to human social cognition? *Mind & Language*, 23 (2), 190-223. [10.1111/j.1468-0017.2007.00337.x](https://doi.org/10.1111/j.1468-0017.2007.00337.x)
- (2010). Intentionality. *Stanford Encyclopedia of Philosophy*. <http://plato.stanford.edu/entries/intentionality/#9>
- (2013). How from action-mirroring to intention-ascription? *Consciousness and Cognition*, 22 (3), 1132-1141. [10.1016/j.concog.2013.02.005](https://doi.org/10.1016/j.concog.2013.02.005)
- (2015). Millikan's teleosemantics and communicative agency. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1-22). Frankfurt a. M., GER: MIND Group.
- Lackey, J. (2005). Testimony and the infant/child objection. *Philosophical Studies*, 162 (2), 163-190. [10.1007/s11098-004-7798-x](https://doi.org/10.1007/s11098-004-7798-x)
- (2006). It takes two to tango: Beyond reductionism and non-reductionism in the epistemology of testimony. In J. Lackey & E. Sosa (Eds.) *The epistemology of testimony* (pp. 160-189). Oxford, UK: Oxford University Press.
- Mascaro, O. & Sperber, D. (2009). The moral, epistemic, and mindreading components of children's vigilance towards deception. *Cognition*, 112 (3), 367-380. [10.1016/j.cognition.2009.05.012](https://doi.org/10.1016/j.cognition.2009.05.012)
- Millikan, R. G. (1984). *Language, thought and other biological categories*. Cambridge, MA: MIT Press.
- (1989). Biosemantics. *The Journal of Philosophy*, 86 (6), 281-297.
- (2004). *Varieties of meaning*. Cambridge, MA: MIT Press.
- (2005). *Language: A biological model*. Oxford, UK: Oxford University Press.
- Sperber, D. (2001). An evolutionary perspective on testimony and argumentation. *Philosophical Topics*, 29 (1/2), 401-413. [10.5840/philtopics2001291/215](https://doi.org/10.5840/philtopics2001291/215)
- Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origg, G. & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, 25 (4), 359-39. [10.1111/j.1468-0017.2010.01394.x](https://doi.org/10.1111/j.1468-0017.2010.01394.x)
- Sperber, D. & Wilson, D. (1986). *Relevance, communication and cognition*. Cambridge, MA: Harvard University Press.
- Van Eemeren, F. H., Meuffels, v. B. & Verburg, M. (2000). The (un)reasonableness of ad hominem fallacies. *Journal of Language and Social Psychology*, 19 (4), 416-435. [10.1177/0261927X00019004002](https://doi.org/10.1177/0261927X00019004002)
- Van Eemeren, F. H., Garrsen, B. & Meuffels, B. (2008). Reasonableness in confrontation: Empirical evidence concerning the assessment of ad hominem fallacies. In F. H. Van Eemeren & B. Garrsen (Eds.) *Controversy and confrontation: Relating controversy analysis with argumentation theory* (pp. 181-195). Amsterdam, NL: John Benjamins.
- Walton, D. (1998). *Ad hominem arguments*. Tuscaloosa, AL: The University of Alabama Press.
- (2008). *Informal logic: A pragmatic approach*. Cambridge, UK: Cambridge University Press.
- Yarp, A. (2013). Ad hominem fallacies, bias and testimony. *Argumentation*, 27 (2), 97-109. [10.1007/s10503-011-9260-5](https://doi.org/10.1007/s10503-011-9260-5)

Assessing a Speaker's Reliability Falls Short of Providing an Argument

A Reply to Marius F. Jung

Pierre Jacob

When confronted with a speaker's assertion, her addressee can either fulfill the speaker's informative intention and accept the new belief or not. If he does, he can either accept the new belief on the sole basis of the speaker's authority or not. If not, then the addressee can examine the reliability of the speaker's assertion. If he does, then he can either check the content of the speaker's assertion with the contents of his own beliefs or scrutinize the speaker herself as the source of the novel information. If the latter, then he can either examine the speaker's epistemic competence in the relevant domain of discourse or the speaker's moral benevolence (or both). None of the above processes amounts to the addressee producing an argument, let alone an *ad hominem* argument. Only if the speaker offers an argument to back her assertion could the addressee commit an *ad hominem* counter-argument in his attempt at rebutting the speaker's.

Keywords

Argument | Assessment of the reliability of a speaker's assertion | Authority | Benevolence | Competence | Fulfilling the speaker's informative intention | Knowledge

In my paper, I probed the gap between the Gricean approach and Millikan's approach to human communicative agency. In particular, I argued in favor of the Gricean separability thesis, i.e., the thesis that the process whereby an addressee fulfills an agent's communicative intention (by understanding or recognizing her informative intention) is distinct from the process whereby the addressee further fulfills (if and when he does) the agent's informative intention (by accepting either a new belief or a new desire for action). I am grateful to Marius F. Jung for his valuable comments on my paper, in which he tries to offer positive suggestions to-

wards bridging the gap between the Gricean separability thesis and (social) epistemology.

In particular, I agree with Marius F. Jung that the issues of whether and to what extent a communicative agent's testimony should or can be assessed as reliable and justified, and thereby construed as knowledge (and not as mere opinion) by her recipient, are of fundamental importance. I also agree with him that it is worthwhile to try and bridge the gap between the psychological investigation of the process whereby an addressee assesses the reliability of a speaker's testimony and the major divide between the reductionist and the anti-reduction-

Author

Pierre Jacob

jacob@ehess.fr

Institute Jean Nicod
Paris, France

Commentator

Marius F. Jung

mjung02@students.uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

ist perspectives in the epistemology of testimony. However, I still want to resist using the particular bridge (or bridges) Jung is building for me. In the following, I want to briefly explain why.

First of all, let us be clear that what we are dealing with here is the addressee's basic epistemic task of assessing the reliability of a communicative agent's (the speaker's) *testimony* or *assertion*, i.e., utterances with truth-conditional contents, because only assertions can be assessed for their reliability or believability. Only a speaker's assertions, not a speaker's requests, can directly enlarge her addressee's knowledge of the world. For the purpose of the discussion of Jung's epistemological project, we should simply ignore addressees' responses to speakers' utterances of requests, i.e., of utterances that lack truth-conditional contents. (I ignore here the fact that a speaker's request may enlarge an addressee's knowledge of the speaker's own character traits.)

Secondly, as I understand it, Jung would like to directly link the investigation of the addressee's task of assessing the reliability of a speaker's assertion to the dispute between the reductionist and the anti-reductionist perspective in the epistemology of testimony. I will reconstruct Jung's basic strategy by means of the six following assumptions.

- He construes the addressee's overall process of assessment of the reliability of a speaker's assertion as an *argument*.
- As I understand it, he also accepts Sperber et al.'s (2010) view that the overall process whereby an addressee assesses the reliability of a speaker's assertion can be divided into two component processes: the assessment of the *authority* of the speaker (who is the *source* of the testimony) and the assessment of the *content* of the speaker's assertion.
- He further focuses on the addressee's assessment of the *authority* of the speaker as the source of the testimony, at the expense of the assessment of the content of the speaker's assertion.

- He links the addressee's assessment of the authority of the speaker as the source of the testimony to *ad hominem arguments*.
- He draws a distinction between local and global *ad hominem* arguments.
- Finally, he argues that only *local*, not global, *ad hominem* arguments are valid methods whereby an addressee can assess the reliability of the speaker's assertion.

I want mainly to take issue with Jung's very first assumption: when assessing a speaker's assertion, the addressee is evaluating the reliability or believability of her utterance. He is *not* arguing with her and therefore not producing an *ad hominem argument*. (Construing the addressee's process of appraisal as an *attack* against the speaker seems far-fetched to me.) In accordance with Jung's second assumption (at least, on my reconstruction of his train of thought), the addressee's appraisal can in turn be seen as a two-fold process: the addressee can focus on either the *content* or the *source* of the speaker's utterance (or both). If the former, then the addressee's task can be construed as a *consistency* check: he checks the compatibility of the truth of the speaker's assertion with the truths of a relevant sub-set of his own beliefs. In the latter case, he scrutinizes some of the speaker's relevant moral or "personal" properties (to use Jung's own phrase). In particular, he will assess the personal authority of the speaker along two main dimensions: her epistemic competence (or knowledge) about the relevant domain of discourse and her moral honesty, i.e., her benevolence towards him.

Of course, the addressee's assessment of the speaker's reliability along these two dimensions is an inferential process, which builds on the addressee's beliefs about both the content of the speaker's assertion and the speaker's personal authority. In an informal sense, it is a reasoning process. But I want to resist the view that this process should be construed as an *argument*, let alone as an *ad hominem* argument. As Sperber et al. (2010) and Mercier & Sperber (2011) have interestingly argued (no pun intended), to *argue* is to try and cause an addressee to accept a new belief (to endorse the truth of

some proposition), by providing explicit *reasons* for it, i.e., by construing it as the *conclusion* of a set of premises from which it derives either deductively or inductively. In fact, arguments are devices used by a speaker in order to try to *overcome* her addressee's reluctance to fulfill her informative intention (i.e., his reluctance to accept a new belief in accordance with her informative intention), on the *sole* grounds of her authority. If so, then speakers (communicative agents) argue, but an addressee doesn't: an addressee *evaluates* the speaker's argument. Of course, an addressee who disagrees with a speaker's argument in favor of some proposition P can turn into a speaker and offer *counter-arguments* to try to cause his opponent to *change her mind* about the truth of P.

1 Conclusion

When a speaker makes an assertion, she commits herself to the truth of some proposition. She thereby knowingly takes the risk that her addressee examines the reliability of her assertion by either checking the content of the asserted proposition or by scrutinizing her epistemic and moral authority. The addressee's choice is to either fulfill the speaker's communicative intention or not. She can further do it on the sole ground of the speaker's authority or not. As I see it, the issue of whether the addressee could *wrong* the speaker by committing some epistemic injustice towards her cannot arise in the process whereby the addressee assesses the reliability of the speaker's mere assertion of P. It can only arise if and when the speaker offers some explicit *argument* in favor of proposition P, in the reasoning process whereby the addressee evaluates the speaker's explicit argument in favor of P, i.e., the link between P and the premises selected by the speaker to justify P. Only then could the addressee produce an *ad hominem* counter-argument (either local or global) meant to successfully or unsuccessfully rebut the speaker's argument for P.

References

- Mercier, H. & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, 34 (2), 57-111.
[10.1017/S0140525X10000968](https://doi.org/10.1017/S0140525X10000968)
- Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G. & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, 25 (4), 359-393.
[10.1111/j.1468-0017.2010.01394.x](https://doi.org/10.1111/j.1468-0017.2010.01394.x)

Wild Systems Theory as a 21st Century Coherence Framework for Cognitive Science

J. Scott Jordan & Brian Day

The present paper examines the historical choice points the led twentieth-century cognitive science to its current commitment to correspondence approaches to reality and truth. Such a “correspondence”-driven approach to reality and truth stands in contrast to coherence-driven approaches, which were prominent in the 1800s and early 1900s. Coherence approaches refused to begin the conversation regarding reality with the assumption that the important thing about it was its independence of observers because the reality-observer split inherent in correspondence-driven views often led to objective-subjective divides, which, within scientific theorizing, tended to render the latter causally unnecessary and in need of ontological justification. The present paper fleshes out the differences between coherence- and correspondence-driven approaches to reality and truth, proposes an explanation of why cognitive science came to favor correspondence approaches, describes problems that have arisen in cognitive science because of its commitment to correspondence theorizing, and proposes an alternative framework (i.e., Wild Systems Theory—WST) that is inspired by a coherence approach to reality and truth, yet is entirely consistent with science.

Keywords

Affordances | Coherence approach to reality and truth | Energy-transformation system | Epistemic gap | Evolutionary theory | External grounding | External relations | Global groundedness thesis | Internal relations | Intrinsic properties | Modes of experience | Realism | Reality | Relational properties | Representational | Self-sustaining embodiment | Ultra grounding | Wild systems theory

1 Introduction

Over the course of its history, cognitive science has often assumed that the important question regarding reality was its independence of an observer. Within this framework, epistemology becomes paramount as scientists work to discover the lawful connections between observer-independent reality and observers. Implicit, if not ex-

plicit, in this approach to cognitive science is the assumption that “truth” is to be measured in terms of the degree of discrepancy between observer-independent reality and whatever impressions, thoughts, representations, affordances, and other observer-dependent phenomena observers use to overcome this assumed epistemic gap.

Authors

J. Scott Jordan

jsjorda@ilstu.edu

Illinois State University

Bloomington-Normal, IL, U.S.A.

Brian Day

bmday15@gmail.com

Clemson University

Clemson, SC, U.S.A.

Commentator

Saskia K. Nagel

s.k.nagel@utwente.nl

University of Twente

Enschede, Netherlands

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität

Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University

Melbourne, Australia

In contrast to such correspondence-driven approaches to reality and truth, many coherence-driving philosophers of the late 1800s and early 1900s rejected correspondence as a starting point for ontology because they believed the subject-object divide it engendered ultimately made it difficult to defend the reality of the subjective (Gardner 2007; Hegel 1971; Priest 1991; Tseng 2003). Given their commitment to the reality of phenomena such as consciousness, value, and meaning, coherence theorists refused to accept the ontological risks inherent in correspondence approaches to reality. Instead, they proposed an alternative approach that admits the reality of consciousness, value, and meaning and assesses truth in terms of the degree of coherence (i.e., non-contradiction) (Oakeshott 1933; Tseng 2003).

In what follows, we flesh out the differences between coherence- and correspondence-driven approaches to reality, propose an explanation of why cognitive science came to favor the correspondence approach, describe problems that have arisen in cognitive science because of its commitment to correspondence theorizing, and propose an alternative framework (i.e., Wild Systems Theory—WST) which is inspired by a coherence approach to reality yet is entirely consistent with science.

2 Correspondence and coherence

2.1 A creation myth: The origins of the correspondence view

A professor walks into the first day of his graduate-level Learning and Cognition course. He tells the students the following story:

“A boy is riding his bike and sees a bracelet on the street. He stops his bike, picks up the bracelet, and realizes the bracelet is a snake.”

After reading the story, the professor asks the students to describe it using the concept “real.” The students share perplexed glances, as if to say, “I signed up for a science course, not a philosophy course.” The professor continues to press the issue, and eventually a student speaks.

“He thought the object was a bracelet, but it was really a snake.”

This prompts another student to say, “He misperceived the snake as a bracelet.”

The professor asks the class if they understand these statements and if they agree with the students’ use of the concept “real.” The vast majority of the class nods yes.

The professor then asks the following, “Is there anything real about the bracelet?”

Eyes roll and students laugh as the question comes across as being silly more than important. The professor waits patiently and asks the question again.

After some time, a student states, “He really believed he saw a bracelet.”

When the professor asks the class if they understand and agree with the statement, only half or less nods yes.

To cut to the chase, the professor asks, “How many of you had a dream in the last week?”

Surprised by the question, few students raise their hand.

Needing to get the class on-board, the professor pushes harder and asks, “Ok. How many of you have had a dream in the past year?”

Now everybody raises their hand.

“Good,” says the professor. “And was there anything real about the dream?”

Connecting the questions regarding the reality of the bracelet and the reality of dreams, a student says, “The dream was real in the sense that I had the experience.”

“Excellent,” states the professor. “Now you understand the type of thinking that lies at the root of our thinking about reality and truth.”

Students look back at him, slightly puzzled.

“According to what you just told me,” the professor begins, “both the snake and the bracelet are real.”

The class continues to stare.

“How many of you think the two are equally real?”

More staring.

“OK. How many of you think the snake is more real than the bracelet?”

Roughly two-thirds of the class raises their hand.

“Why?”

One student raises her hand and states, “The boy really experienced a bracelet, but since the bracelet was an incorrect perception, the snake is more real.”

“And when the boy finally had a snake perception,” states the professor, “his perception was correct?”

“Yes,” responds the student confidently.

“Excellent!” exclaims the professor. “How many agree?”

The students look back and forth to each other, seeking an answer. Eventually, most everyone in the class raises their hand.

“Now we are truly making progress,” states the professor, “and for my final question, how do we know the snake is more real than the bracelet?”

The same student answers without hesitation, “Because the snake perception accurately corresponds to the object.”

“There it is,” exclaims the professor. “We know the object is really a snake because our experiences correspond to it. In short, perceptions are true, or accurate, because they correspond to reality correctly.”

He centers himself in front of the class and states, “This way of describing reality is known as the correspondence approach to truth and reality. It has dominated the way we think about truth and reality for at least four hundred years, if not longer. And over the next two weeks I hope to show you that if you believe this approach to truth and reality, you, one, logically deny yourself access to reality, and, two, make it very difficult to defend the reality of phenomena such as love, hate, the sound of music, and the taste of ice cream.”

He looks out over the class and sees that he has their attention.

“How many of you really like ice cream?” he asks.

Everyone raises their hand instantly. Some students raise both hands.

“Good then,” the professor states. “Let us begin.”

2.2 A very brief history of correspondence, reality, and truth

While the story described above may seem rudimentary, the purpose is to give the reader, as well as the hypothetical student, a common entry point into the conversation regarding correspondence and coherence approaches to reality and truth. This is important because coherence approaches have not been proposed all that often over the past one hundred years. Thus, very few contemporary cognitive scientists know of them, let alone make use of them. This century-long waxing and waning of correspondence and coherence approaches, respectively, may have had something to do with the fact that alternatives to correspondence have come to be seen as increasingly irrelevant after a century of naturalism, physicalism, and realism. That is, the increasingly sophisticated view of the physical world that has developed over centuries of scientific practice has led the vast majority of practicing cognitive scientists to assume that the issue of reality and truth has been solved, and by using science, we decrease the degree of discrepancy between objective and subjective reality. From this perspective, science is metaphysical in the sense that science reveals how reality really is, independent of our personal perspective.

While this correspondence-driven, metaphysical take on science is practically implicit in contemporary cognitive science, we propose that the issues addressed in the snake/bracelet story are, in fact, unresolved. Furthermore, we believe that the current zeitgeist of correspondence thinking is due to historical choices regarding our conceptualization of the reality of human experience. In what follows, we briefly review some of these choice points in the hope of clarifying why a commitment to correspondence has seemed to be such an obvious step for cognitive scientists.

a. Spiritual versus mental subjectivity. Questions about whether or not the bracelet is real, or the manner in which it is real in relation to the reality of the snake, are the same kind of questions René Descartes asked himself when he addressed the reality of God and the

material world hundreds of years ago. To be sure, very few if any contemporary cognitive scientists would account for the reality of the snake and the bracelet via Descartes's notion of interacting yet qualitatively distinct physical and spiritual realities (i.e., dualism). However, despite their assumed distinctiveness from dualism, most contemporary cognitive scientists implicitly, if not explicitly, endorse the basic assumption of dualism that the interesting point about reality is the extent to which it is independent of observers. This commitment to correspondence thinking was evident in the writing of one of Descartes' major critics, [John Locke \(1700\)](#). Even after Locke took some of the first formal steps toward developing cognitive science (i.e., a "science of man") and re-described the spiritual side of Descartes' dualism as being "mental," the question for Locke's "science of man" was how it is that our sense impressions are able to accurately correspond to physical reality.

b. Radical skepticism. In response to Locke's non-spiritual correspondence approach to reality and truth, [David Hume \(2012\)](#) asked whether or not such an approach is even logically possible. Specifically, Hume's basic argument was that if one accounts for reality in terms of the "impressions and ideas" it causes within us, then all we can ever really know are the impressions and ideas we have about reality. This is because every test we could ever run to assess the extent to which our impressions and ideas about external reality are accurate would have to be mediated by impressions and thoughts. That is, once we claim that we know external reality through observer-dependent structures such as thoughts and impressions, we have logically doomed all of our knowledge to be trapped within us.

Though Hume's radical skepticism is hundreds of years old, and seems outdated to many contemporary scientists in general—and cognitive scientists, specifically—we believe Hume's radical skepticism constitutes both a historical choice point and an individual choice point for the issue of how we conceptualize the reality of the subjective. On the one hand, there were and are those scholars who took radical skepticism

to be diagnostic of a logically flawed approach to reality and truth. On the other, there were and are those who believed and continue to believe that the test for whether or not the correspondence approach to reality and truth is "correct" is empirical. That is, the "correctness" of science will ultimately be decided on correspondence grounds; that is, by whether or not science can eventually represent the entirety of observer-independent reality accurately. In what follows, we examine various historical attempts to sustain the correspondence approach in spite of radical skepticism.

c. Overcoming radical skepticism.

What is somewhat ironic about the attempt to overcome skepticism is that although those who did and do so tend to present themselves as being quite different from each other, they nonetheless avoid skepticism in roughly the same way; specifically, by nesting the correspondence relation within an assumed, larger-scale reality that guarantees the veridicality of the correspondence relation. Descartes, for example, after having doubted all but his ability to doubt, then went on to infer that his ability to do so could have only been created by a superior, omnipotent being (i.e., God). Then, to secure the correspondence relationship completely, he assumed that his subjectivity must correspond accurately to reality because God created both and would not have done so incorrectly. Bishop Berkeley made much the same maneuver when he proposed to overcome Hume's radical skepticism by asserting that the correspondence relation holds because we exist within God's mind.

Cognitive scientists, while certainly not dualists, nonetheless rely on evolutionary theory as a means of placing the correspondence relationship within a larger-scale reality as a means of validating the correspondence relationship. There are two dominant varieties of such thinking: indirect-realism and direct-realism. Realism is the assertion that objects exist as they are, with all of their intrinsic properties, independently of observers. Indirect realism asserts that our knowledge of reality is mediated by our sensory systems and knowledge structures. Direct realism asserts that our knowledge struc-

tures are directly in contact with external properties, exactly as they are.

Indirect realism is basically an evolutionarily inspired re-description of Locke's mediated theory of perception in which external events cause the internal formation of impressions and ideas. Though there are many varieties of indirect realism (Fodor 1983; Pinker 1999), common to most is the computationalist, representationalist view of cognition, which assumes that we know what is outside of us because of the representations that external events cause within our brains. Given that our brains co-evolved with the world and were naturally selected, it seems self-evident that our brains give us accurate access to external reality.

While in the early days of cognitive science indirect realists believed that internal, sensory-driven (i.e., bottom-up) representation of external events could be augmented by top-down, cognitive processes such as attention (Broadbent 1958; Cherry 1953), they still nonetheless believed that the bottom-up processes entailed accurate representations of their external causes. Such assumptions derived support from findings such as Hubel & Wiesel's (1962) discovery of neurons in the primary visual cortex (V1), that expressed spatially correspondent receptive fields (i.e., the activity of a neuron in V1 could be maximally stimulated by a visual stimulus emanating from a particular location in the visual field). Later research revealed a massive degree of spatial correspondence between locations in external space and neural space within a host of different modalities (e.g., visual, auditory, and kinesthetic space). Milner & Goodale's (1995) discovery of visual systems used for object identification versus visual systems used for guiding action (i.e., vision for perception versus vision for action) further solidified indirect realism because it seemed to clarify how internal representations of external events were used to accurately guide behaviors back onto external reality.

In light of this accumulating neural evidence as well as a host of perceptual-cognitive research that revealed our apparent ability to represent invariant properties of biologically relevant external events, Roger Shepard (2001)

stated the following in the opening line of the abstract to his seminal paper, *Perceptual-cognitive Universals as Reflections of the World*: "The universality, invariance, and elegance of principles governing the universe may be reflected in principles of the minds that evolved in that universe" (p. 581). Clearly, from this indirect-realist perspective, our connection to the world around us is mediated by internal representations that are phylogenetically derived stand-ins for what the world around us is like.

Critiques of indirect realism within cognitive science basically recapitulated Hume's critique of Locke's mediated theory of perception. That is, cognitive scientists dating back as far as the Six Realists (Holt et al. 1910) criticized the representational approach to cognition because they believed it logically denied one access to external reality. Interestingly enough, instead of challenging the correspondence view of reality and truth that lay at the heart of indirect realism, and which constituted Hume's biggest concern with Locke's approach, cognitive scientists who labeled themselves direct-realists argued that the connections between the internal and the external were not constituted of mediating representations of the external but, rather, of natural relations between the organism and the environment. Though this idea dates back at least as far as William James as well as the Gestalt psychologist Kurt Koffka (Ash 1998), perhaps its most influential expression was provided by J. J. Gibson (1979), who argued that we perceive the world in terms of behavioral possibilities, what he referred to as affordances.

Since Gibson (1979), many cognitive scientists have effectively investigated affordances. Given that most ecological psychologists who investigate affordances are simultaneously direct realists, it is important to their realism that affordances be real, and that we have direct access to affordances via our sensory systems. Instead of constructing representations, however, our sensory systems are described as having the task of picking up or detecting information (i.e., affordances).

The direct-realist appeal to the reality of directly perceivable affordances defends the

validity of the correspondence relation by arguing that organisms veridically perceive affordances because they evolved to do so. That is, just as was the case with Descartes, Berkeley, and indirect realism, the assertion of the correspondence relation is validated by placing it within an *assumed*, larger-scale reality. In the case of direct realism, that assumed, larger-scale reality is the evolved physical world.

By calling the *evolved, physical world* an assumed, larger-scale reality, we are not proposing that the theory of evolution is untrue, or that the phenomena referred to via the concept of the physical world do not exist. In fact, we believe the phenomena referred to via the concept of the physical world do exist, and we further believe that the theory of evolution is “true.” We just believe they exist and are true respectively, in a manner that is not couched in the correspondence framework espoused by realists. (We will describe how we believe they exist and are true at a later point in this paper.) Rather, what we are trying to accomplish by referring to the evolved physical world as an assumed, larger-scale reality is to point out the common strategy shared by correspondence theorists across the centuries. Specifically, if one espouses a correspondence account of reality, in which knowledge and/or perceptual structures are meant to correspond to reality, either via perceptually generated representations or via evolutionarily tailored relations, then, by definition, all we have contact with are knowledge and perceptual structures, and any statement about external reality is an *assumption*. This, in fact, was the gist of Hume’s critique of Locke’s mediated theory of perception. Radical skepticism does not argue that objects do not exist. Rather, it is simply a critique of a *particular* account of reality (i.e., the correspondence account), and the critique refers to the *logical coherence* of the account. If one espouses a correspondence framework for reality and truth, one has logically denied oneself access to external reality, and neither empirical data nor an assumed larger-scale reality is capable of overcoming this logical flaw. On logical grounds alone, one cannot use realism and its attendant correspondence arguments to overcome radical skepticism.

To be sure, direct realists might respond that their brand of realism overcomes radical skepticism because direct realism does not rely on internal representations to connect the internal to the external. Rather, the connections, as stated above, are conceptualized in terms of relations between organisms and environments that co-evolved in such a way that organisms are able to directly perceive these relations (i.e., affordances).

While at first glance the anti-representational slant of these arguments does seem to skirt the issue of radical skepticism, it’s appeal to relations or relational properties between relata (e.g., organisms and environments) still commits to the correspondence notion that truth is determined by the degree of correspondence between the system (i.e., the organism) and something external to the system (i.e., affordances). Again, this commitment to the correspondence relation stems from the centuries-old belief that the important thing about reality is its independence of observers. Armed with such an approach to reality and truth, science is believed to be metaphysical in that it reveals observer-independent properties of external reality. To be sure, the direct realist will argue that evolution has solved all of this. However, as was stated above, it is their commitment to realism that logically denies the correspondence scholar access to external reality. In short, it is the logically incoherent notion of correspondence that denies the realist access to external reality, not reality itself.

2.3 The coherence approach to reality and truth

In order to overcome the representationalism inherent in indirect realism, direct realists re-framed the connection between organisms and environments in terms of evolutionarily derived relations as opposed to internal representations. Doing so, however, begs the issue of the nature of the things that stand in relation to each other (i.e., the relata). Are the relata themselves constituted of relational properties? If so, just how far down is reality constituted of relations?

While questions regarding the relational nature of reality might seem silly to contemporary cognitive scientists, it was actually of paramount importance to the maintenance and perpetuation of the correspondence approach roughly a century ago. [Bertrand Russell \(1911\)](#), for example, went to great lengths to counter the notion of internal relations that was prominent in idealist philosophy in the 1800s and early 1900s. As described by Russell, the notion of internal relations is the idea that the relations between entity A and B are actually constituents of A and B. In other words, part of what constitutes A is its relationship with B. This idea was problematic for Russell because idealist philosophers often used it as a means of overcoming radical skepticism. Specifically, these philosophers proposed that the objectivity of supposed external reality was actually observer dependent, in that a subject (i.e., an observer) was internally related to its objects. That is, the objects do not have an existence independent of the subject, and vice versa ([Hegel 1971](#); [Oakeshott 1933](#); [Priest 1991](#)). Different idealist philosophers held different motivations for espousing this view. Many did so in order to maintain the reality of God. Others did so in order to maintain the reality of phenomena that Descartes had relegated to the subjective (e.g., values, meaning, and aesthetics).

Regardless of their motivations, the idealist notion of internal relations was problematic for Russell because he wanted to describe reality in terms of the objects of science and logic. In short, Russell wanted metaphysics to be empirical. In order to do so, he felt he needed to establish the logical independence of external reality. That is, he had to show that objects are not internally related to subjects. As a result, he argued that not all relations are internal, and that some are external. By external relations, Russell meant that a relationship between entity A and entity B is not constitutive of entities A and B. An example of an external relation would be the relative height of two people, say Mary and Sam. While it is logically coherent to state that Mary is taller than Sam, the “taller” relation is not constitutive of either Mary or Sam. That is, the “taller” relation de-

pends, of course, upon Mary and Sam, but it exists externally from Mary and Sam in the sense that it plays no role in the properties that constitute Mary or Sam. Russell uses this notion of external relations to propose a correspondence approach to reality and truth in which entities share relations and via those relations constitute components of complexes. Having assumed that he had logically negated the notion of internal relations, Russell then proposed that we get on with the empirical, metaphysical business of scientifically describing reality “as it is,” independent of observers.

The use of the notion of externally related entities as a means of sustaining the correspondence approach to reality and truth is also evident in the work of direct realists such as [Holt et al. \(1910\)](#) and [Gibson \(1979\)](#). By utilizing this relation-driven form of realism, all three were implicitly asserting the belief that the issue of reality was to be solved via epistemology. That is, they were continuing the centuries-old argument that the important thing about reality is its independence from observers.

a. The relational nature of reality. As stated above, the direct-realist assumption that we have contact with external reality via relations begs the issue of the nature of the things that stand in relation to each other (i.e., the *relata*). In other words, if we claim that two *relata* share a relation, we imply that there is a difference between *relata* and relations. This leads to another choice point that historically influenced the manner in which we describe the reality of the subjective: Are the *relata* themselves constituted of relational properties, or are they constituted of non-relational properties, what one might refer to as *intrinsic* properties? The answer to this question is important, for if one argues for a difference between intrinsic and relational properties, then realism seems the obvious choice; the purpose of science is to uncover the intrinsic properties of reality. If, however, one assumes that *relata* are themselves constituted of relational properties, we have a much different problem. For if all *relata* are constituted of relations, then there can be no intrinsic properties. This is because the constitution of all properties, by definition, would be re-

lational. In short, reality would constitute a unity in which all things were constituted of all things.

The notion that all things are about all things sounds much like the idealist notion of internal relations. And while the idea might seem outdated in contemporary cognitive science, it has recently gained traction in the philosophy of science as a possible explanation of properties. For example, mass is often considered an intrinsic property in that the mass of an object is considered to be independent of its context, while weight is considered to be an extrinsic property because the object's weight is determined by how its mass interacts with its context. Jammer (2000), however, proposes that all particles receive their inertial mass via their interactions with the Higgs field, "a scalar field that 'permeates all of space' and 'endows particles with mass'" (p. 162). Bauer (2011) argues that the dependence of mass on the Higgs field renders mass *externally grounded*. This means that the mass of the particle is not independent of its context. As a result, the object's mass is a *relational, non-intrinsic* property.

Bauer's notion of external grounding should not be confused with Russell's (1911) notion of external relations. Bauer uses the notion of external grounding to make the case that a property (i.e., mass) that was assumed to be intrinsic (in order to distinguish it from the property of weight, which was assumed to be contextually relative) was actually contextually relative. "External" in this sense was used to flesh out the relative nature of a previously assumed to be non-relative (i.e., intrinsic) property (i.e., mass). Russell, on the other hand, used the concept "external" in the opposite way. That is, he wanted to demonstrate that certain properties were independent (i.e., were not entailed in the constitution) of other properties. In short, Russell used the notion "external" to create independent properties in a reality the idealists had described as an internally related unity, while Bauer, roughly a century later, uses the concept "external" to re-contextualize properties that post-Russellian realists had conceptually isolated from reality by describing them as *intrinsic*.

While one could see Russell's (1911) and Bauer's (2011) uses of the concept "external" as contradictory and leave it at that, one might also argue that their different uses of the same concept are diagnostic of the success of Russell's efforts. Specifically, Russell used the concept external to de-contextualize certain parts of reality (i.e., make them intrinsic), while Bauer, one hundred years later, uses the same concept to re-contextualize what Russell had worked so hard to de-contextualize. In short, one might argue that while Russell represented a first conceptual step away from holism, contemporary works such as Bauer's represent initial conceptual steps back toward holism. Further evidence of a tendency to conceptually move the philosophy of science away from the notion of intrinsic properties can be found in the work of Harré (1986), who proposes the notion of *ultra-grounding*, the idea that a property may be grounded by a property, or properties, of reality as a whole.

Such an anti-intrinsic take on the nature of properties is also proposed by both Schaffer (2003) and Dehmelt (1989). These authors assert that there may be no fundamental level to reality at all (i.e., no final, non-relational, intrinsic property that forms "relations" into "complexes"). Rather, they propose that reality may be constituted of infinite levels of microstructure. Consistent with the notion of external grounding, Prior et al. (1982) propose the Global Groundedness Thesis. This thesis asserts that all dispositions (i.e., properties) are grounded (i.e., externally grounded) rather than ungrounded (i.e., intrinsically grounded). Ladyman et al. (2007) implicitly, if not explicitly, express a similar critique of the notion of intrinsic properties when they assert that contemporary analytic metaphysics needs to abandon the idea that reality is constituted of self-subsistent individual objects.

b. Truth in a relational reality. The idea that reality is infinitely relational is inconsistent with the correspondence approach to reality and truth because a relational reality can never be subdivided into final, intrinsic, "in-and-of-themselves"-type properties. In an infinitely relational reality, all objects

and subjects are composed of relations (i.e., they are contextually grounded), and all intrinsic properties are inherently relational. This implies that dialectic counterparts such as objective versus subjective, or relational versus intrinsic, come to be introduced into one's description of reality, not because they reflect accurate, final, ontological subdivisions of reality, but for the same reason one describes the snake in the snake-bracelet story as being more real than the bracelet—specifically, because one accepts the subjective-objective divide inherent in the correspondence view and tries to defend the assumed greater reality of the snake by asserting its independence of oneself. It is this assumption that the important thing about reality is its assumed observer-independent nature that drives the correspondence approach and leads one to further believe that the goal of science is to overcome subjectivity and reveal the objective truth about reality. Once such independence is no longer assumed, then truth can no longer be measured by assessing the degree of difference between reality and an impression, idea, or representation we have of it, or by investigating an assumed relation we share with it. There exists nothing “as it is” to which anything else can accurately correspond. The final, ontological description of what something *is* must include reality as a whole. In short, truth must be assessed in a non-correspondence fashion.

One way to measure truth without asserting a correspondence relationship is to do so on the basis of coherence. By coherence we mean lack of contradiction. In contemporary philosophy, lack of contradiction (i.e., coherence) is most often used to refer to the means by which a belief is justified (Kvanvig 1995; Lycan 2012). Specifically, a subset of contemporary epistemologists, who might be loosely referred to as “coherentists” (Lycan 2012; Quine & Ullian 1978; Thagard 1978), propose a view akin to the following:

[...]what justifies [...] the formation of any new belief—is that the doxastic move in question improves the subject's explanat-

ory position overall and/or increases the explanatory coherence of the subject's global set of beliefs. (Lycan 2012, p. 6)

While the coherentist approach to propositions clearly relies on the notion of “lack of contradiction” to measure the justifiability of beliefs, it does not make use of “lack of contradiction” as a measure of the truth inherent in experience. As a result, it is logically possible for one to be a coherentist about beliefs while simultaneously holding an implicit or explicit correspondence view that conceptualizes beliefs as subjective propositions that refer to external, objective reality. It is not clear where Lycan (2012) stands on this issue.

At any given moment, we find ourselves involuntarily holding any number of beliefs, at least those produced by perception and by memory; however, [...] I do not make any primary appeal to those faculties as justifying. Call such unconsidered beliefs “spontaneous beliefs”; they are primarily about our immediate environment, past events, sometimes our own mental states, and more. (p. 6)

Although Lycan (2012) makes no claims regarding the metaphysical status of *perception*, or where he stands on the issue of reality and experience, his use of the word *perception* allows him to interject other phrases such as “primarily about our immediate environment,” that then implicitly connect beliefs to external reality via a correspondence relation. Regardless of whether or not this was Lycan's intent, it is clear that coherentism is about the justifiability of beliefs and not about reality, per se. As a result, it may not have much to offer in our attempt to develop a coherence approach to reality and experience.

One possible way to apply the coherence approach to the issue of reality and experience is the very same test entailed in the snake-bracelet problem. If one assumes that reality constitutes an internally related unity that defies that logic of correspondence tests of truth, then statements regarding the truth of the

snake and the bracelet should be stated in terms of contradiction. That is, the statement “the boy saw a bracelet while riding his bike” is true in the sense that the boy had a persistent flow of “bracelet” experience. The notion of persistent flow is important here because it calls attention to the fact that from moment to moment during the bracelet phenomenon, the phenomenon did not contradict itself; that is, the “bracelet” phenomenon at one moment was not followed by a different “non-bracelet” phenomenon the next. Jordan & Vandervert (1999) propose that it is this coherent flow of phenomena, what they refer to as “within-instance” coherence, that underlies our propositions regarding the reality of phenomena. To be sure, later on in the story, when the boy picked up the “bracelet,” he suddenly did have a contradiction in the flow of the bracelet phenomenon; specifically, the bracelet phenomenon was contradicted by a “snake” phenomenon. Given that the snake phenomenon persisted in a more coherent fashion than the bracelet phenomenon (i.e., no matter what he did, the boy could not convert the snake phenomenon into another type of phenomenon), one then asserts that the snake phenomenon is more real than the bracelet phenomenon. From the coherence perspective, what this means is that the snake phenomenon was more coherent (i.e., more persistent, or less contradictory) than the bracelet phenomenon.

Such a coherence approach to the reality and truth of phenomena is rather similar to the approach advocated by Michael Oakeshott. In perhaps his most famous book, *Experience and its Modes*, Oakeshott (1933) described reality in a manner that is consistent with the idea that reality constitutes an internally related unity. He did not say it this way, however. Rather, as was consistent with both his idealist background and the philosophical context of his time, he described reality in terms of experience and stated, “[...]experience is a single whole, within which modification may be distinguished, but which admits of no final or absolute division” (Oakeshott 1933, p. 27). Also,

[s]ubject and object are not independent elements or portions of experience; they

are aspects of experience which, when separated from one another, degenerate into abstractions. Every experience [...] is the unity of these, a unity which may be analysed into these two sides but which can never be reduced to a mere relation between them. (Oakeshott 1933, p. 60)

To be sure, the manner in which Oakeshott uses the concept of experience makes it difficult for those who have already made correspondence-driven commitments to the meaning of “experience” to follow his arguments. For correspondence theorists, “experience” refers to the subjective side of Descartes dualism. But given that Oakeshott did not define experience in terms of the mental, spiritual, transcendental, or absolute, it seems reasonable to assume that when he described reality as a world of experience, he was using the concept differently than it had been used by Locke, Kant, or Hegel. This is important, for when most contemporary cognitive scientists refer to idealism, they tend to mention Locke and Berkeley (Charles 2011). Locke and Berkeley both accepted the correspondence relation. Locke accepted it without reservation. Berkeley accepted it and then placed it within the assumed larger-scale reality of God’s mind in order to avoid skepticism. Oakeshott, on the other hand, denied the correspondence relation (as did most all the German idealist philosophers). Thus, for Oakeshott, the terms “reality” and “experience” were synonymous, not because he believed reality was ultimately subjective, but because he believed reality constituted an internally related unity that defied any ontological, final division into dialectic categories such as subjective and objective, or reality versus experience.

c. Coherence, truth, and modes.

Oakeshott proposed his coherence approach to reality and truth because he believed that the correspondence approach was, first, logically incoherent, and second, improperly applied in contexts in which it was not relevant. Specifically, Oakeshott argued that within the confines of the correspondence approach, it was easy to believe that the task of science was to uncover the intrinsic, observer-independent properties of

reality. In addition, given its supposed ability to accumulate a stockpile of context-independent, universal knowledge, it became easy to believe that its criterion for truth (i.e., correspondence) should have dominion over all arenas in which truth was at stake.

Agreeing with his idealist predecessors about the logical incoherence of correspondence thinking, Oakeshott argued that endeavors such as “science” constituted modes of experience. What he meant by “mode” is that science constitutes a distinct means of generating abstractions about the internally related unity in which we are embedded. It is an abstraction in the sense that it is constitutive of reality (i.e., it is “within” the reality it is attempting to describe) and can therefore never be “outside” of reality, looking “at” reality. As a result, it should be conceptualized as a recursion on reality—an abstraction about that from which it emerged and within which it is entailed.

Oakeshott described at least four different modes: science, daily practice (i.e., politics), history, and poetry. What distinguishes these modes, in addition to the content they are about, is the means by which truth is determined within each. In the mode of science, truth is determined by the degree of quantitative coherence that can be achieved in the description of a phenomenon, both individually and collectively. Given that quantitative coherence within and between individuals is paramount, factors such as personal opinion are irrelevant to the truth criteria of the mode of science. In the mode of daily practice (i.e., politics), however, opinion and desire (i.e., how people want to live their lives) constitute the issue at hand. Truth, therefore, could not be measured in terms of the degree of quantitative coherence within and between individuals. Rather, it was reflected in the degree to which members of a group treated each other in accordance with a normatively determined system of expectations. As a result, the truth criteria of the modes of science and politics (i.e., daily practice) were similar in that they were both measured in terms of coherence but were fundamentally different in terms of the phenomena whose coherence was being as-

sessed (i.e., quantification of a phenomenon versus normatively determined expectations).

Because of this qualitative difference in the relation of science and politics, Oakeshott argued that the truth criteria of one could not coherently be used to measure the truth of the other. That is, just as personal opinion and desire were to play no role in the truth status of scientific statements, quantitative coherence in both individual and collective descriptions should not play a role in determining the truth status of political statements (i.e., statements of how people should live their lives).

Oakeshott went to such great lengths to distinguish science as a mode of experience because he felt he needed to provide an alternative to the correspondence approach. By appealing to the notions of *coherence* and *internally related unity* that were common to idealist philosophers, without making appeals to the mental, spiritual, transcendental, or absolute, Oakeshott presented a coherence approach that was capable of addressing the physicalist, naturalist forms of correspondence thinking that were emerging during his time. The difference between Oakeshott’s coherence approach and the correspondence-driven naturalism of his time was not that the former did not believe in the reality of objects or that the former was created to maintain a place for God in metaphysics, as had been the case for Berkeley and Kant. Rather, the difference was that the former recognized the logical incoherence of the latter and worked to develop an approach to reality that avoided the logical pitfalls historically encountered by the latter. Given that direct realists such as [Holt et al. \(1910\)](#) and [Gibson \(1979\)](#), who were, to some extent, contemporaries of Oakeshott, had probably developed fairly robust associations between coherence, idealism, and the religious agendas of Berkeley and Kant, they probably had no reason to assume that an idealist-inspired philosophy had anything to offer.

Regardless of who did or did not read Oakeshott’s work while he was alive, his lack of appeal to mental, spiritual, transcendental, or absolutist themes, coupled with his persistent attacks on the correspondence approach, collectively support the idea that when he referred to

reality as a world of experience, he was using it more as a placeholder in his arguments with the correspondence approach as a way to slowly transform the reader's meaning of the word experience from the subjective-mental denotation it had acquired in the midst of the correspondence approach to the holist-driven, internally related unity of all phenomena it was meant to imply in the coherence framework.

d. Coherence and science. To correspondence ears, the description of the coherence approach given above might be interpreted as antiscientific. That is, since we take seriously the logical incoherence of the correspondence approach and assert that it does not inform us about context-independent, intrinsic properties of reality, one might assume we are proposing that science does not reveal truth. This is a common reaction of those who implicitly hold a correspondence view. They assume that those who acknowledge the strength of Hume's insight are actually denying the existence of "things." This is simply not the case. As stated above, radical skepticism is a critique of the internal logic of the correspondence approach to reality and truth, not a critique of the existence of "things." Oakeshott's coherence approach constitutes a means of addressing reality and truth in a way that does not beg incoherent correspondence assumptions. In order to further demonstrate the compatibility of science and the coherence approach, we present WST as a case in point. As we present WST we will also point out how various choice points in the theory's construction were guided by the notion of coherence.

3 Wild systems theory

WST is a recently developed theory of cognitive systems (Jordan 2008, 2013; Jordan & Ghin 2006, 2007; Jordan & Heidenreich 2010; Jordan & Vinson 2012) that conceptualizes organisms in a different light than technological metaphors such as switchboards and computers, or dynamical metaphors such as Watt Governors and convection rolls. Rather, WST follows the lead of physicists (Schrödinger 1992), theoretical biologists (Kauffman 1995) and ecologists (Odum

1988), and conceptualizes organisms as multi-scale, self-sustaining energy-transformation systems. What is meant by *self-sustaining* is that the *work* of the system (i.e., the energy exchanges that actually constitute the system, such as the chemical work that constitutes biological systems) gives rise to products (e.g., other chemicals) that serve as a catalyst for the reaction that produces the product or some other reaction in the system. When a self-catalyzing system of work emerges, it is able to *sustain* itself as long as the proper fuel source remains available.

What is meant by *multi-scale* is that an organism can be coherently conceptualized as being constituted of different scales of self-sustaining work. Jordan & Vinson (2012) describe the notion of multi-scale, self-sustaining work in the following manner:

At the chemical level, self-sustaining work has been referred to as autocatalysis (Kauffman 1995), the idea being that a self-sustaining chemical system is one in which reactions produce either their own catalysts or catalysts for some other reaction in the system. At the biological level, self-sustaining work has been referred to as autopoiesis (Maturana & Varela 1980), again, the idea being that a single cell constitutes a multi-scale system of work in which lower-scale chemical processes give rise to the larger biological whole of the cell which, in turn, provides a context in which the lower-scale work sustains itself and the whole it gives rise to (Jordan & Ghin 2006). Hebb (1949) referred to the self-sustaining nature of neural networks as the 'cell assembly', the idea being that neurons that fire together wire together. Jordan & Heidenreich (2010) recently cast this idea in terms of self-sustaining work by examining data that indicate the generation of action potentials increases nuclear transcription processes in neurons which, in turn, fosters synapse formation. At the behavioural level, Skinner (1976) referred to the self-sustaining nature of behaviour as operant conditioning, the idea being

that behaviours sustain themselves in one's behavioural repertoire as a function of the consequences they generate. [Streeck & Jordan \(2009\)](#) recently described communication as a dynamical self-sustaining system in which multi-scale events such as postural alignment, gesture, gaze, and speech produce outcomes that sustain an ongoing interaction. And finally, [Odum \(1988\)](#) and [Vandervert \(1995\)](#) used the notion of self-sustaining work to refer to ecologies in general. (p. 235)

3.1 Wild systems theory and coherence

Conceptualizing organisms as being composed of multi-scale, self-sustaining work is consistent with coherentism ([Lycan 2012](#)). That is, the notion of self-sustaining work increases the coherence of our conceptualization of organisms (i.e., beliefs about organisms) because it reveals the dynamic homologies that transcend both the phyla and the nesting of multi-scale, energy-transformation systems that constitute a single organism. From plants, to neurons, to behavior, to persons, to human societies, increasingly complex systems of work (i.e., energy transformation) have evolved precisely because the work of which they are constituted is self-sustaining in that the work produces catalysts for either the work itself or some other level of work in the multi-scale system.

When we conceptualize organisms with technical metaphors such as switchboards and computers, we leave out these homologous, multi-scale, energy-transformation dynamics that living systems do not have in common with technological systems. This use of technological metaphors then forces us to generate explanations of the means by which our technologically inspired model of the organism is “connected” to the external context. To be sure, the issue is not unique to science. Descartes ran into the same problem when he divided humans into physical and spiritual substrates, and most scholars who have taken Descartes's correspondence problem seriously have had to do something similar. Locke proposed causal connections between external events and internal im-

pressions and ideas. Kant proposed a priori conceptions of space and time. Indirect realism proposed evolutionarily derived representations, and direct realism proposed evolutionarily derived “relations.”

Given its focus on multi-scale, self-sustaining homologies, WST is able to focus on that which is common across the internal and external contexts of an organism; namely, energy transformation. As a result, WST's focus on internal/external homologies renders it consistent with the coherence approach to reality and truth. Specifically, its focus on internal/external homologies prevents WST from internal/external conceptualizations that lead to the connection problems experienced by correspondence-driven approaches. Within contemporary correspondence frameworks (e.g., indirect and direct realism), the external context tends to be conceptualized as physical. Historically, the concept physical has garnered its meaning from its dialectic relationship with concepts such as “mental” and “spiritual.” As a result, its usage implicitly intimates a correspondence relation and leaves us having to determine whether or not the internal context is likewise physical, mental, or something altogether different, as well as how it is that the internal context is connected to the external context.

Within WST, the internal and external contexts of an organism are both conceptualized in terms of energy transformation. Specifically, the external context is conceptualized as a self-organizing, energy-transformation hierarchy ([Odum 1988](#); [Vandervert 1995](#)), while brains and organisms are conceptualized as multi-scale, self-sustaining energy transformation systems that are able to sustain themselves in the larger-scale energy transformation hierarchy because the work of which they are constituted produces its own catalysts. Inspired by this idea, [Jordan & Ghin \(2006\)](#) proposed that *the fuel source dictates the consumer*. This means that any system that sustains itself on a certain fuel source (e.g., plants on sunlight, herbivores on plants, or carnivores on herbivores) must be constituted such that it is able to address the constraints involved in capturing that fuel source.

Conceptualizing organisms as self-sustaining embodiments of the contextual constraints entailed within an energy-transformation hierarchy renders WST consistent with a coherence approach to reality and truth because an embodiment of context is necessarily “about” that context. By “necessarily” we mean that the system’s internal dynamics are phylogenetically and ontogenetically emergent from the energy-transformation hierarchy in which it sustains itself; it is an embodiment of the reality (i.e., context) within which it emerged. In short, it is reality within reality. The idea that organisms constitute embodiments of context is consistent with [Friston’s \(2011\)](#) assertion that organisms constitute an embodiment of an optimal model of their environment. Interestingly enough, Friston is led to this assertion for much the same reason WST is led to its notion of organisms as embodied contexts; specifically, because both begin with the idea of the organism as an energy-transformation system. As a result,

there is no epistemic gap between an organism and its environment. Organisms do not need to be ‘informed’ by environments in order to be about environments because they are necessarily ‘about’ the contexts they embody. Rather, what self-sustaining systems need do is sustain relationships with the contexts in which they are embedded in ways that lead them to sustainment. According to WST, meaning is constitutive of embodied context (i.e., bodies). As a result, living systems are necessarily meaningful ([Jordan, 2000a](#)), not because a body is alive or dead, because it is physical, or because it is biological. Living is meaning because it is sustained, embodied context. ([Jordan & Vinson 2012](#), p. 9)

Given this lack of an epistemic gap between embodiments of context and the contexts in which they sustain themselves, WST dissolves the subjective-objective epistemic barrier created by the correspondence approach. Embodiments of context are naturally and necessarily “about” their context and, as a result, are inherently meaningful.

Our use of the word *meaningful* is not meant to imply that the evolutionary emergence of living systems simultaneously constituted the emergence of meaning into a reality that had been, up until then, *meaningless*. Rather, our equating the notion of *embodied context* with *meaningfulness* is meant to demonstrate the serious metaphysical consequences that emerge from our earlier description of reality as an internally related unity. If all phenomena are, in the end, contextually dependent, then part of what constitutes them is their relation with the rest of reality. In short, as was stated previously, self-sustaining systems are reality within reality. It is this irreducible, inherent relationality that we are conceptualizing as *meaning*.

Within contemporary philosophy of mind, it might seem as though we are asserting that embodied contexts (i.e., self-sustaining bodies) *instantiate* phenomenal properties. While this assertion is not incorrect, our concern with such an interpretation is the implicit, correspondence-driven assumption that phenomenal properties are subjective while other properties of the system are objective. Our take on this issue is that embodied contexts do not represent the emergence of phenomenology into reality as much as they represent the emergence self-sustaining relationality into reality. And it is this self-sustaining relationality that phylogenetically scales up to the phenomenon we refer to via terms such as consciousness and phenomenology.

Defining meaning in this way allows for meaning (i.e., embodied context) to be constitutive of what organisms are. As a result, phenomena traditionally referred to via concepts such as phenomenology, consciousness, meaning, and value, which tended to be relegated to the subjective/internal side of correspondence frameworks and had to be described as being emergent from, identical with, or fundamentally different from “physical” properties ([Chalmers 1996](#)), are considered phylogenetically scaled-up versions of the embodied meaning inherent in all embodied contexts. [Jordan & Vinson \(2012\)](#) describe why it is that self-sustaining embodiments of context entail meaning:

In a single-cell organism, the internal dynamics (i.e., the micro scale) and the organism as a whole (the macro scale) are coupled in such a way that changes in the micro-scale (e.g., low energy levels) give rise to changes at the macro-scale (e.g., behaviors such as swimming and tumbling) that recursively influence the micro-scale (i.e., give rise to energy intake) and, in the end, foster the sustainment of both levels of scale. In short, the micro-macro coupling is self-sustaining. In the case of a rock, the micro-macro coupling is not recursively self-sustaining. The coupling generates no dynamics that serve to sustain a particular aspect of either the macro or micro organization. (Jordan & Vinson 2012, pp. 11-12)

Jordan & Ghin (2006) refer to the embodied aboutness of a single-cell organism as *proto-consciousness*. They do so for the following reasons: (1) to acknowledge the meaning (i.e., embodied context) inherent in a single-cell (i.e., a small-scale, self-sustaining embodiment of context), and (2) to set the groundwork for an explanation of how the proto-consciousness of a single-cell system could possibly scale up to the full-blown self-awareness entailed in humans. As regards this scaling up, Jordan & Vinson (2012) say the following:

It was possible for self-sustaining systems to scale-up from the level of single-cell organisms to the level of human beings because their status as energy-transformation systems simultaneously rendered them a potential fuel source for any system that embodied the constraints necessary to sustain itself on such embodied energy. As an example, the emergence of herbivores gave rise to a context that afforded the emergence of carnivores. A significant constraint of being a carnivore, however, was the need to capture a moving fuel source. Doing so required, and still requires, anticipatory structures regarding the future location of the moving target. Jordan and Ghin (2006) assert that the embodiment of

anticipatory dynamics in the neuromuscular architecture of organisms capable of propelling themselves as a whole toward anticipated locations constituted the phylogenetic emergence of anticipatory aboutness. That is, the self-sustaining dynamics of one system came to be ‘about’ the future dynamics of another system. WST equates such anticipatory aboutness with the traditional notion of mind, and proposes that phenomena that have received labels such as memory, thought, phenomenology, and self-awareness constitute evolutionary recursions (i.e., scale-ups) of the anticipatory dynamics embodied in self-sustaining systems. Given that all self-sustaining systems constitute embodiments of context and are, therefore, necessarily ‘about’ context, their anticipatory dynamics likewise entail ‘aboutness.’ Thus, as self-sustaining systems evolved and became increasingly abstract (i.e., about increasingly abstract events such as tomorrow, next week, and/or next year), meaning, too, became increasingly abstract. (Jordan & Vinson 2012, p. 12)

WST’s conceptualization of meaning as embodied context is consistent with Oakeshott’s (1933) coherence approach to reality and truth in that it does not assume that subjects and objects are independent and in need of connection. Rather, subjects (i.e., organisms) are considered embodiments of their context and are, therefore, internally related to their context. The contexts in which they are and have been embedded are constitutive of what they are. Said in a more familiar way, a thoroughgoing (i.e., maximally coherent), ontologically minded explanation of what an organism *is* must include all aspects of the organism as well as the contexts it embodies.

To be sure, WST is not the only approach to propose that (1) organisms constitute embodiments of their contexts, and (2) such systems necessarily entail anticipatory dynamics. As was stated previously, Friston (2011) makes a similar claim when he asserts that (1) organisms constitute optimal models of their environ-

ments, and (2) they utilize anticipatory coding as a means of optimally maintaining homeostasis. (See Andy Clark's, Jakob Hohwy's, and Anil Seth's contributions to this collection for other approaches to cognition that posit a reliance on anticipatory coding.). A potential difference between WST and Friston's position is the degree of metaphysical commitment WST makes to the assertion that reality constitutes an internally related unity. That is, while Friston's view is consistent with the notion of embodied contexts, it is not clear he also agrees with the coherence approach to reality. As a matter of fact, much of his explanation of how it is that organisms generate and maintain minimum free energy is couched in the epistemic language of external stimuli and internal representations. Though the use of these terms does not, in and of itself, indicate a commitment to direct or indirect realism, it does reveal, at the very least, a minimal, implicit commitment to a correspondence approach to reality and truth.

This comment on Friston's position should not be construed as a critique of his framework, as much as it should be taken to constitute a means by which the unique metaphysical commitments of WST can be thrown into sharp relief. Friston's goal is to provide a maximally coherent account of the causality underlying cognition. The goal of WST is to provide a scientifically informed approach to reality and truth that does not rely on the correspondence relation. The difference in these missions fairly thoroughly accounts for the differences between WST and Friston's free energy approach, and the jury can still be out as to whether or not the free-energy principle constitutes a correspondence approach to reality and truth.

3.2 Wild systems theory and truth

As was stated previously, a coherence approach to reality and truth assesses the degree of truth in experience and beliefs via the degree of coherence entailed in and across both. As was also previously stated, this coherence approach to truth differs from coherentism (Lycan 2012) in that the latter applies the criterion of coher-

ence (i.e., lack of contradiction) to beliefs, while the former applies it to both experience (i.e., moment-to-moment contradictions in experience) and beliefs.

Given this notion of the organism as a self-sustaining prediction, WST is able to apply the coherence criterion to both experience and beliefs because it conceptualizes organisms as embodiments of context and avoids the correspondence relation. As a result, truth is not measured in terms of the degree of correspondence between the subjective and the objective. Rather, it is measured in terms of the degree of non-contradiction entailed within one's moment-to-moment embodied context (i.e., phenomenology) and across the beliefs one derives from the moment-to-moment flows of embodied context. In Friston's (2011) language, the degree of coherence in an embodied context might be taken to refer to the degree of prediction error minimization that has been achieved by the organism's current model of reality. To make this work however, and to avoid the implicit epistemic gap implied by the notion of a "model of reality," the meaning of the word *model* would have to be stretched to such a point that the organism itself constitutes a model of reality. To be sure, Friston intimates as much when he describes organisms as optimal models of their environments. To make this use of the word *model* simultaneously imply that the organism-as-model *constitutes* anticipation, the organism itself would have to be seen as constituting a prediction. While this use of the concept prediction seems strange, it is actually consistent with how Friston uses the term when describing the chemotactic behaviors exhibited by *E. coli*:

...by selective modulation of tumbling frequency, these bacteria show chemotaxis. This is a nice example of an itinerant policy based on the prior expectation (endowed by natural selection) that the organism will only change its motion through state-space when it encounters unexpected (costly) generalized states (here, a decrease in the concentration of attractants). (2011, p. 114)

What is at issue here is the degree of ontological commitment entailed in Friston's assertion that natural selection endows organisms with prior expectations. Is he claiming that organisms are constituted of phylogenetically derived prior expectations, or is he simply presenting prior expectations as a productive way to model organisms? While his assertion that organisms constitute optimal models of their environments seems to favor the former interpretation, his later use of terms such as sensations and representations seems to favor the latter. Whatever the case, if Friston's notion of minimizing prediction error is to be used as a description of what it means for there to be a contradiction in the flow of contingent context, then the concept *prediction* has to be used in a way that does not engender an epistemic gap. In short, the organism has to be conceptualized as a self-sustaining prediction.

In order to better clarify this admittedly abstract means of talking about truth, we offer certain arguments presented in the present paper as a case in point. As was mentioned previously, indirect- and direct-realist approaches to reality and experience rely on evolutionary theory as a means of connecting the subjective and the objective. In our critique of these views, we argued that they validated the correspondence relation by conceptually placing it within the assumed, larger-scale reality of the evolved physical world. WST, however, also makes use of an assumed, larger-scale reality, specifically, the self-organizing, energy-transformation hierarchy (Odum 1988). The difference between the two uses of evolutionary theory lies in what the two approaches are believed to reveal about evolution. To realists, be they direct or indirect realists, evolutionary theory is believed to reveal reality as it is, independent of observers. Within WST, evolutionary theory is definitely seen as being "true," but in the coherence sense that it is the most coherent account of the existence of species yet given.

When describing the "truth" of evolutionary theory in coherence terms, it is important to remember that WST is not radically skeptical about whether or not the phenomena referred to via the realist notion of an *evolved physical world* (e.g., organisms, rocks, and plants) exist. To the contrary, it would be incoherent to deny our belief

that such phenomena exist and do so outside of our skin. What is at stake is the issue of *how* something exists beyond our skin. In a correspondence framework, what is important about something existing on the other side of our skin is that it be observer-independent. Given this conceptualization, one has to explain how observer-independent and observer-dependent phenomena are connected. In the coherence framework, the existence of objects beyond the skin, as well as the idea that they exist as such without the presence of an observer, is conceded. However, defining their reality status in terms of their observer-independence is seen as being insufficient, for even though they may exist independently of the presence of an *observer*, such observer-independence in no way implies such objects exist independently of all context. No phenomenon, no matter how universal, exists as it does independently of all other phenomena. In short, all phenomena are context-dependent.

WST's notion of embodied context implies that we should measure the truth status of claims made in cognitive science in terms of their degree of coherence, both within experience and across beliefs. Given that most contemporary cognitive scientists are direct or indirect realists, either explicitly or implicitly, they tend to assume the correspondence relationship (again, either explicitly or implicitly), which, in turn, makes it difficult for them to coherently address the reality of "subjective" phenomena such as phenomenology, meaning, and value. To be sure, by aligning itself with a coherence approach to truth, WST logically denies itself access to objective, intrinsic reality. But given that WST conceptualizes the notion of objective, intrinsic reality as an incoherent assumption derived from the coherence of moment-to-moment experience, WST, simply given its commitment to coherence, could not accept such a notion in the first place.

3.3 Wild systems theory and cognitive science

Given that WST is not designed to reveal intrinsic properties of objective reality, its beliefs about science are inconsistent with the correspondence notion that science is metaphysical.

Let us recall that the slogan “science is metaphysical”, which was briefly mentioned at the beginning of the present paper, is just shorthand for the philosophical thesis that the goal of science is to overcome the objective-subjective divide and reveal the “real,” observer-independent, intrinsic properties of reality. By asserting that all properties are contextually grounded and cannot therefore be intrinsic, WST posits that science cannot reveal intrinsic properties. As a result, there is no final, thing-as-it-is essence to which any “experience” or “theory” can correspond. As a further result, there can be no correspondence test of reality. Science, therefore, cannot be metaphysical. This lack of belief in the metaphysical nature of science, however, is in no way anti-scientific. On the contrary, it is wholly consistent with Oakeshott’s (1933) contention that the practice of science constitutes a mode of experience. That is, if reality is an internally related unity, then theories are constitutive of that reality and can never “point to” reality as if to do so outside of it. They are, by definition, “in it” just as we are. Thus, they are, by definition, incomplete, what Oakeshott referred to as an arrestment of the whole (i.e., a mode of experience). As an example, WST’s scientifically inspired conceptualization of organisms as self-sustaining embodiments of context does assume a “larger-scale reality” within which organisms are nested, just as direct and indirect realism do. The different reasons for doing so are important. In WST, a larger-scale reality is assumed because it would be incoherent not to do so. That is, we would be contradicting both our experiences and our beliefs about those experiences if we claimed we did not exist within something larger than ourselves. From the correspondence perspective, a larger-scale reality is assumed, and it is believed to comprise observer-independent, intrinsic properties that science will ultimately reveal.

An immediate implication of coherence-versus correspondence-driven approaches to science is that while the latter conceptualizes science as inherently metaphysical (i.e., it reliably reveals intrinsic, observer-independent properties of objective reality), the former conceptual-

izes science as a method by which we are able to increase the coherence of our statements about that within which we are embedded (i.e., coherentism; Lycan 2012). Such coherentism is valuable because it affords us more influence over our context; that is, it affords us the ability to more effectively sustain ourselves.

To be sure, the idea that the value of science is pragmatic, as opposed to metaphysical, is not new. Dewey (1929) proposed much the same:

But the search does not signify a quest for reality in contrast with experience of the unreal and phenomenal. It signifies a search for those relations upon which the occurrence of real qualities and values depends, by means of which we can regulate their occurrence. To call existences as they are directly and qualitatively experienced ‘phenomena’ is not to assign to them a metaphysical status. It is to indicate that they set the problem of ascertaining the relations of interaction upon which their occurrence depends. (Dewey 1929, pp. 103–104)

Interestingly enough, Dewey espoused his pragmatic approach to science for much the same reason Oakeshott proposed his coherence approach to reality and truth—specifically, because they both believed that the realist, physicalist naturalism of their time was inspired by a logically incoherent correspondence framework that had been historically derived from dualism’s assumed split between spiritual and material reality. Dewey states,

The notion that the findings of science are a disclosure of the inherent properties of the ultimate real, of existence at large, is a survival of the older metaphysics. It is because of injection of an irrelevant philosophy into interpretation of the conclusions of science that the latter are thought to eliminate qualities and values from nature. This created the standing problem of modern philosophy:— the relation of science to the things we prize and love and

which have authority in the direction of conduct. (1929, p. 102)

As regards cognitive science specifically, WST's coherence approach to the meaning of science provides a way for cognitive scientists to experience their theories and models as pragmatic tools versus metaphysical tests. In addition, WST's reliance on the concept of embodied context provides a means for cognitive scientists to discuss those phenomena traditionally associated with the subjective side of correspondence theorizing (e.g., phenomenology, value, and meaning) without relying on the subjective-objective correspondence relation. This is important, for as was mentioned in the latter half of the preceding quote by Dewey, by conceptualizing the practice of science as a means of overcoming the correspondence relationship, realist philosophers ultimately put the reality of the "subjective" at risk as more and more naturalists came to conceptualize the subjective in terms of inherently meaningless, physical properties (Gardner 2007). As was stated previously, by conceptualizing organisms as self-sustaining embodiments of context, WST renders properties that had been historically associated with the subjective, such as phenomenology, value, and meaning (see Jordan & Vinson 2012, for a thorough review of this issue), constitutive of what organisms are. As a result, cognitive scientists can avoid distracting arguments about such correspondence-driven issues as the grounding problem (i.e., how do concepts and symbols garner their meaning; Harnad 1990), or the relationship between the physical brain and consciousness. These issues are only experienced as important, hard problems within the conceptual confines of correspondence theory and the belief that the answer will be found via cognitive science.

4 Conclusions

To be sure, there were twentieth-century philosophers other than Dewey and Oakeshott whose approach to reality and truth was very consistent with the coherence approach. Heidegger and Merleau-Ponty are two examples. Perhaps these rela-

tionships will be fleshed out to a greater extent in future papers. For the present paper, the purpose was to (1) illustrate for the reader that there is another, historically relevant, robust approach to reality and truth other than the correspondence approach, and (2) illustrate that this other approach is completely consistent with science.

Maybe it was the fact that many idealist philosophers used their anti-correspondence frameworks as a means of defending the reality of God that led so many scientifically minded philosophers to avoid it to the point that now, after more than one hundred years of neglect, it is rarely if ever mentioned or utilized in cognitive science. This is precisely why we began this paper with the snake-bracelet story. Coherence approaches have been out of fashion for so long that we felt it necessary for the reader to experience, first hand, the type of thinking that has always fostered questions about reality. Our assumption was that by experiencing the tension between what it means to describe the snake as real and what it means to describe the bracelet as real, the readers would be in a better position to understand that although the coherence approach was ignored during the past century, Oakeshott's presentation of a non-spiritual, non-absolute, non-transcendental coherence framework leaves the coherence and correspondence frameworks on similar, logical ground. Given the advent of concepts such as external grounding, ultra grounding, and *global groundness* in contemporary philosophy of science, it seems the coherence approach to reality and truth is, at the very least, once more being discussed.

Wild Systems Theory is only one possible theory of "what people are" that could emerge from a coherence-driven perspective, and we suspect there will be others. But given WST's description of phenomenology as an evolutionarily, scaled-up form of self-sustaining embodied context, phenomena such as the taste of ice cream are rendered just as "real" as the cream and sugar that constitute the ice cream. We believe this is an important achievement. And when one considers WST's compatibility with science, it seems reasonable to propose WST as a twenty-first-century coherence framework for cognitive science.

References

- Ash, M. G. (Ed.) (1998). *Gestalt psychology in German culture, 1890-1967: Holism and the quest for objectivity*. Cambridge, UK: Cambridge University Press.
- Bauer, W. (2011). An argument for the extrinsic grounding of mass. *Erkenntnis*, 74 (1), 81-99. [10.1007/s10670-010-9269-4](https://doi.org/10.1007/s10670-010-9269-4)
- Broadbent, D. E. (1958). *Perception and communication*. New York, NY: Pergamon Press.
- Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford, UK: Oxford University Press.
- Charles, E. P. (Ed.) (2011). *A new look at new realism: The psychology and philosophy of E.B. Holt*. New Brunswick, CN: Transaction Publishers.
- Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, 25 (5), 975-979. [10.1121/1.1907229](https://doi.org/10.1121/1.1907229)
- Dehmelt, H. (1989). Triton, ...Electron..., Cosmon...: An infinite regression? *Proceedings of the National Academy of Sciences*, 86 (22), 8618-8619. [10.1073/pnas.88.4.1590a](https://doi.org/10.1073/pnas.88.4.1590a)
- Dewey, J. (1929). *The quest for certainty: The study of the relation of knowledge and action*. New York, NY: Minton, Balch & Company.
- Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. Cambridge, MA: MIT Press.
- Friston, K. J. (2011). Embodied inference: Or I think therefore I am, if I am what I think. In W. Tschacher & C. Bergomi (Eds.) *The implications of embodiment: Cognition and communication* (pp. 89-125). Exeter, UK: Imprint Academic.
- Gardner, S. (2007). The limits of naturalism and the metaphysics of German idealism. In E. Hammer (Ed.) *German idealism: Contemporary perspectives* (pp. 19-49). Abingdon, UK: Routledge.
- Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin.
- Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42 (1), 335-346. [10.1016/0167-2789\(90\)90087-6](https://doi.org/10.1016/0167-2789(90)90087-6)
- Harré, R. (1986). *Varieties of realism: A rationale for the natural sciences*. Oxford, UK: Blackwell.
- Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. New York, NY: Wiley.
- Hegel, G. W. (1971). *Hegel's philosophy of mind*. Oxford, UK: Clarendon Press.
- Holt, E. B., Marvin, W. T., Montague, W. P., Perry, R. B., Pitkin, W. B. & Spaulding, E. G. (1910). The program and first platform of six realists. *The Journal of Philosophy, Psychology and Scientific Methods*, 7 (15), 393-401. [10.2307/2010710](https://doi.org/10.2307/2010710)
- Hubel, D. H. & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160 (1), 106-154.
- Hume, D. (2012). *A treatise of human nature*. New York, NY: Courier Dover Publications.
- Jammer, M. (2000). *Concepts of mass in contemporary physics and philosophy*. Princeton, NJ: Princeton University Press.
- Jordan, J. S. (2008). Wild-agency: Nested intentionalities in neuroscience and archeology. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 363 (1499), 1981-1991. [10.1098/rstb.2008.0009](https://doi.org/10.1098/rstb.2008.0009).
- (2013). The wild ways of conscious will: What we do, how we do it, and why it has meaning. *Frontiers in Psychology*, 4 (574), 1-12. [10.3389/fpsyg.2013.00574](https://doi.org/10.3389/fpsyg.2013.00574)
- Jordan, J. S. & Ghin, M. (2006). (Proto-) consciousness as a contextually-emergent property of self-sustaining systems. *Mind & Matter*, 7 (4), 45-68.
- (2007). The role of control in a science of consciousness: Causality, regulation and self-sustainment. *Journal of Consciousness Studies*, 14 (1-2), 177-197.
- Jordan, J. S. & Heidenreich, B. (2010). The intentional nature of self-sustaining systems. *Mind & Matter*, 8 (1), 45-62.
- Jordan, J. S. & Vandervert, L. (1999). Liberal education as a reflection of our assumptions regarding truth and consciousness: Time for an integrative philosophy. In J. S. Jordan (Ed.) *Modeling consciousness across the disciplines* (pp. 307-331). New York, NY: University Press of America.
- Jordan, J. S. & Vinson, D. (2012). After nature: On bodies, consciousness, and causality. *Journal of Consciousness Studies*, 19 (5/6), 229-250.
- Kauffman, S. (1995). *At home in the universe*. New York, NY: Oxford University Press.
- Kvanvig, J. L. (1995). Coherentism: Misconstrual and misapprehension. *Southwest Philosophy Review*, 11 (1), 159-168. [10.5840/swphilreview199511116](https://doi.org/10.5840/swphilreview199511116)
- Ladyman, J., Ross, D., Spurrett, D. & Collier, J. (2007). *Every thing must go: Metaphysics naturalized*. New York, NY: Oxford University Press.
- Locke, J. (1700). *An essay concerning human understanding*. London, UK: Black Swan.

- Lycan, W. G. (2012). Explanationist rebuttals (coherentism defended again). *The Southern Journal of Philosophy*, 50 (1), 5-20. [10.1111/j.2041-6962.2011.00087.x](https://doi.org/10.1111/j.2041-6962.2011.00087.x)
- Maturana, H. & Varela, F. (1980). *Autopoiesis and cognition: The realization of the living*. Boston, MA: Reidel.
- Milner, A. D. & Goodale, M. A. (1995). *The visual brain in action*. Oxford, UK: Oxford University Press.
- Oakeshott, M. (1933). *Experience and its modes*. Cambridge, UK: Cambridge University Press.
- Odum, H. T. (1988). Self-organization, transformity, and information. *Science*, 242 (4882), 1132-1139. [10.1126/science.242.4882.1132](https://doi.org/10.1126/science.242.4882.1132)
- Pinker, S. (1999). How the mind works. *Annals of the New York Academy of Sciences*, 882 (1), 119-127. [10.1111/j.1749-6632.1999.tb08538.x](https://doi.org/10.1111/j.1749-6632.1999.tb08538.x)
- Priest, S. (1991). *Theories of the mind: A compelling investigation into the ideas of leading philosophers on the nature of the mind and its relation to the body*. New York, NY: Mariner Books.
- Prior, E., Pargetter, R. & Jackson, F. (1982). Three theses about dispositions. *American Philosophical Quarterly*, 19 (3), 251-257.
- Quine, W. V. O. & Ullian, J. S. (1978). The web of belief (Vol. 2). In R. M. Ohmann (Ed.) New York: Random House.
- Russell, B. (1911). The basis of realism. *The Journal of Philosophy, Psychology and Scientific Methods*, 8 (6), 158-161.
- Schaffer, J. (2003). Is there a fundamental level? *Noûs*, 37 (3), 498-517. [10.1111/1468-0068.00448](https://doi.org/10.1111/1468-0068.00448)
- Schrödinger, E. (1992). *What is life?: With mind and matter and autobiographical sketches*. Cambridge, UK: Cambridge University Press.
- Shepard, R. N. (2001). Perceptual-cognitive universals as reflections of the world. *Behavioral and Brain Sciences*, 24 (4), 581-601. [10.1017/S0140525X01000012](https://doi.org/10.1017/S0140525X01000012)
- Skinner, B. F. (1976). *About behaviorism*. New York, NY: Vintage Books.
- Streeck, J. & Jordan, J. S. (2009). Communication as a dynamical self-sustaining system: The importance of time-scales and nested contexts. *Communication Theory*, 19 (4), 445-464. [10.1111/j.1468-2885.2009.01351.x](https://doi.org/10.1111/j.1468-2885.2009.01351.x)
- Thagard, P. R. (1978). The best explanation: Criteria for theory choice. *The Journal of Philosophy*, 76-92.
- Tseng, R. (2003). *The sceptical idealist: Michael Oakeshott as a critic of the enlightenment (Vol. 1)*. Charlottesville, VA: Imprint Academic.
- Vandervort, L. (1995). Chaos theory and the evolution of consciousness and mind: A thermodynamic-holographic resolution to the mind-body problem. *New Ideas in Psychology*, 13 (2), 107-127. [10.1016/0732-118X\(94\)00047-7](https://doi.org/10.1016/0732-118X(94)00047-7)

Thickening Descriptions with Views from Pragmatism and Anthropology

A Commentary on J. Scott Jordan & Brian Day

Saskia K. Nagel

How can we as biological systems that are self-organizing and constantly adapting make sense of our surroundings? How can the rich connections between organisms and environment lead to our particular lifeworlds, lifeworlds that allow individual experiences and that are themselves constantly changing in reaction to them? This commentary suggests, extending the framework provided by Scott Jordan and Brian Day, an integration of recent neuroscientific evidence with perspectives from pragmatism, anthropology, and phenomenological thought. Much experimental evidence demonstrates that human beings are systems comprised of a brain as part of a body and an environment, which is constantly regulating and adapting. This evidence resonates with reasoning from pragmatism and anthropology that describe the continuous, dynamic interaction of mind, body, and world. Employing those various perspectives leads to a dense description of human experience and cognition that specifies details and patterns, which considers contextual factors that allow us to enrich human self-understanding, and which aids attempts to answer the questions raised at the beginning of this paper.

Keywords

Anthropology | Circular causalities | Enactivism | Mind-body-world-relationship | Pragmatism | Systems approach

Commentator

[Saskia K. Nagel](#)
s.k.nagel@utwente.nl
University of Twente
Enschede, Netherlands

Target Authors

[J. Scott Jordan](#)
jsjorda@ilstu.edu
Illinois State University
Bloomington-Normal, IL, U.S.A

[Brian Day](#)
bmday15@gmail.com
Illinois State University
Bloomington-Normal, IL, U.S.A.

Editors

[Thomas Metzinger](#)
metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

Mind as background is formed out of modifications of the self that have occurred in the process of prior interactions with environment. Its animus is toward further interactions. Since it is formed out of commerce with the world and is set toward that world nothing can be further from the truth than the idea which treats it as something self-con-

tained and self-enclosed. ([Dewey 1934](#), p. 269)

Knowing does not lie in the establishment of a correspondence between the world and its representation, but is rather immanent in the life and consciousness of the knower as it unfolds within the field of practice set up through his or her presence as a being-

in-the-world [...]. Like life itself, the unfolding does not begin here or end there, but is continually going on. It is equivalent to the very movement—the processing—of the whole person, indivisibly body and mind, through the lifeworld. (Ingold 2001, p. 159)

1 Introduction

Philosophers and scientists alike have long been interested in the question of how our being-in-the-world allows us to experience in a plethora of ways and to behave meaningfully. In extending the framework suggested by Scott Jordan and Brian Day, this commentary suggests integrating recent neuroscientific evidence with perspectives from pragmatism, anthropology, and phenomenological thought. The commentary shall be programmatic in the sense that it prepares the way for further argument and discussion by making available new perspectives that invite the reader to look beyond the “classical” argument and thus benefit from various disciplines. The driving questions are: How can we as biological systems that are self-organizing and constantly adapting make sense of our surroundings? How can we grasp our world via perception? How can we skillfully engage with the world? How can the rich connections between organisms and environment lead to our particular lifeworlds; lifeworlds that allow individual experiences and that are themselves constantly changing in reaction to them?

One dominant approach to reality and truth has been the correspondence approach of computational cognitive sciences that assumes that reality can be revealed by science, independently of the personal perspective of an observer. The task of correspondence theories is to understand the relation between observer and observer-independent reality; a task that assumes dichotomies between inner and outer, between objective and subjective. Facing the limits of those approaches, [Scott Jordan & Brian Day \(this collection\)](#) suggest bridging the riff between the inner and the outer by acknowledging that there is in fact no gap between the organism and its environment. If one wants to

avoid the dualistic trap that asks how something inside the “mind”—such as thoughts or ideas—can represent the outside world, one challenges the seemingly essential dependence of cognitive science on representations.

Much neuroscientific, psychological, anthropological, and philosophical work, both old and new, suggests that we understand cognition as arising from the actions of embodied agents that engage skillfully in a meaningful world ([Beauchamp & Martin 2007](#); [Brooks 1991](#); [Clark 1997](#); [Graziano et al. 1994](#); [Lakoff & Johnson 1999](#); [Noë 2004](#); [O'Regan & Noë 2001](#); [Thompson 2010](#); [Varela et al. 1991](#); [Wilson & Knoblich 2005](#)). This understanding can ultimately help us avoid the correspondence theorists' notorious problem, how the external is connected to the internal. Organisms that are embedded and situated do not need to represent the external environment as they are always already about the contexts in which they live. Moreover, for the situated organism, “the situation is organized from the start in terms of human needs and propensities which give the facts meaning, make the facts what they are, so that there is never a question of storing and sorting through an enormous list of meaningless, isolated data” ([Dreyfus 1992](#), p. 262). Understanding organisms as always already existing in meaningful interaction¹ with their environment and thereby constantly adapting and changing is relevant not just for topics in philosophy of mind but also for epistemology and metaphysics. The metaphysical question of how mind, body, and world are related is tightly linked to epistemological questions about how we can experience the external world. The central tenet is how experience can happen at all, i.e., how the experiencing organism can relate meaningfully to the world.

This commentary furthers the line of thought described by [Scott Jordan & Brian Day](#)

¹ Due to lack of a better concept, the term “interaction” will be used throughout this article even though it entails clearly separable entities that have previously been independent—an assumption that is contested by the approach suggested here. Moreover, due to limited space, this commentary cannot take into account the aspect of intersubjectivity. The relevance of others with whom interaction takes place is inherent in the concept of mind and its interdependence with the environment (see e.g., [De Jaegher & di Paolo 2007](#)).

([this collection](#)) by suggesting further perspectives from neuroscience, pragmatism, and anthropology for approaching cognitive systems as experiencing, bodily systems that are in constant, value-laden interaction with the world; rather than as systems that primarily mirror an external reality from a position separated from the world. Here, I will combine arguments from John Dewey, in particular his work on experience, and anthropologist Timothy Ingold, with recent neuroscientific approaches that support a view that challenges classical correspondence approaches. This will allow a thicker description, i.e., a dense description specifying details and patterns and considering contextual factors, of human experience and cognition.

2 Pragmatism and anthropology meet the neurosciences

In line with much neuroscientific work today, Dewey describes how life is about constantly striving for greater adaptation and for a balance of energies. He beautifully elaborates:

Life itself consists of phases in which the organism falls out of step with the march of surrounding things and then recovers unison with it—either through effort or by some happy chance. And, in a growing life, the recovery is never mere return to a prior state, for it is enriched by the state of disparity and resistance through which it has successfully passed. If the gap between organism and environment is too wide, the creature dies. If its activity is not enhanced by the temporary alienation, it merely subsists. Life grows when a temporary falling out is a transition to a more extensive balance of the energies of the organism with those of the conditions under which it lives. ([Dewey 1934](#), p. 535).

This view resonates with Wild Systems theory, as suggested by [Jordan & Day \(this collection\)](#), which explains an organism not as a computational input–output system but as an open energy-transforming system that must absorb, transform, and use energy to sustain itself. This

does not forestall computation, of course, but it describes the computational process in a different context.

The description of this context can be developed further to challenge correspondence theories: correspondence theories suggest that we understand cognition when we understand how humans represent the external world internally, and when we understand how they process this representation. The focus on a potentially disembodied input–output machine that passively receives information about an observer-independent reality and that has an isolated computational system processing representations cannot tell us how the internal relates to the external—the notorious problem of traditional cognitivism—or how the internal can be enacted in real-world situations that are often vague and constantly changing. As Andy Clark explicates:

Real embodied intelligence [...] is fundamentally a means of engaging with the world—of using active strategies that leave much of the information out in the world, and cannily using iterated, real-time sequences of body-world interactions to solve problems in a robust and flexible way. The image here is of two coupled complex systems (the agent and the environment) whose joint activity solves the problem. In such cases, it may make little sense to speak of one system’s representing the other. ([Clark 1997](#), p. 98)

Cognition and experience arise from ongoing interaction with an unstable, changing environment. The entanglement of the brain, the rest of the body, and its particular environment—which includes other organisms—is essential for experience and reason. This is not the trivial claim that the brain cannot exist without a body; even though the bodily context is often neglected in research studying brain processes.²

2 The importance of the body was put forward by Maurice Merleau-Ponty in the *Phenomenology of perception*: “[t]he body”, he wrote, “is the vehicle of being in the world, and having a body is, for a living creature, to be involved in a definite environment, to identify oneself with certain projects and be” “continually committed to them” (1962, p. 82), and further: “[o]ur bodily experience of movement is not a particular case of knowledge; it provides us *with a way of access to the world* and the object, with a ‘praktognosia’, which has to be recognized as original and

The message is that reason, cognition, mind arise from this very entanglement. How the body relates to the environment structures experiences; there is an immediate coupling between perception and action. Cognition is not a transcendent aspect detached from “matter” (the brain and the rest of the body in particular) but is constantly shaped, fostered, and constrained by the environment and the body’s peculiarities.

Anthropologist Timothy Ingold consequently questions whether it makes sense:

to attribute that quality of the operation of a cognitive device [...] which is somehow inside the animal and which, from its privileged site, processes the data of perception and pulls the strings of action. Indeed it makes no more sense to speak of cognition as the functioning of such a device than it does to speak of locomotion as the product of an internal motor mechanism analogous to the engine of a car. Like locomotion, cognition is the accomplishment of the whole animal, it is not accomplished by a mechanism interior to the animal and for which it serves as a vehicle. (Ingold 1993, p. 431)

It is thus the interaction of the different systems that is the most fascinating research topic in cognitive science—a topic that requires a holistic approach. Such reasoning that considers circular causalities can be traced back to earlier thinkers such as Bateson 1973, Kelso 1995, Maturana & Varela 1980, Thompson 2010, Varela 1996 or von Uexküll 1940. This idea of circular causality as a property of living, self-organizing systems refers to the connection of perception and movement that underlies the ongoing co-constitution of organism and environment. There is continuous top-down-bottom-up interaction that captures the interrelations between several levels in a hierarchy. The gen-

eral underlying idea is that individual small-scale parts enable the existence of order parameters that in turn determine the behavior of the individual parts. Thomas Fuchs (2012) refers to physicist Hermann Haken’s 2004’s work on synergetics, the science of self-organization, to further illustrate the mutually-constraining relation between the microscopic and macroscopic elements of a complex system. Dynamic system modeling in various fields relies on multi-level causal processes in which higher-order processes are mutually entrained with lower-order processes, without one taking precedence over the other (Engel et al. 2001; Freeman 1995; Lewis 2005; Thelen & Smith 1994).

While a purely cognitivist approach that fosters “The Myth of the Inner; The Myth of the Hidden; and The Myth of the Single” (Torrance 2009, p. 112) is still fairly mainstream, in recent years we have seen a growing interest on the part of cognitive scientists and neuroscientists in particular in the relevance of the complex interplay of brain, body, and world. Today, this interplay is finally considered in the empirical study of cognition, which resonates in the growing body of work in cognitive science.³ The importance of embodiment is widely appreciated in cognitive science today. There is a large body of evidence from the neurosciences on how an ongoing organism–environment interaction is essential for cognition (Beauchamp & Martin 2007; Brooks 1991; Chiel & Beer 1997; Engel et al. 2001, 2013). While we still see attempts to describe what has been termed the “‘filing cabinet’ view of mind: the image of the mind as a storehouse of passive language-like symbols waiting to be retrieved and manipulated by a kind of neural central processing unit” (Clark 1997, p. 67)—there is growing consensus that cognition can best be studied and understood in dynamic, interactionist terms, as bound to bodily organisms that are confronted with particular problems in specific environments.

perhaps as primary. My body has its world, without having to make use of ‘symbolic’ or ‘objectifying function’” (1962, p. 140–141; emphasis mine). This has been elaborated and enriched in the last years with views on recent empirical work by Shaun Gallagher (2005), who offers an account of the body that emphasizes the role of embodied action in perception and cognition.

³ Curiously, there is little direct reference to the pragmatists and in particular to John Dewey’s work. Notable exceptions are Mark Johnson (e.g., 2007) and Jay Schulkin (2009), who offer nuanced and explicit pragmatist views on neuroscientific research. Philip Kitcher (2012) offers a wide and detailed demonstration of the importance of pragmatism for philosophy.

Dewey once again can serve as an inspiring reference point:

To see the organism in nature, the nervous system in the organism, the brain in the nervous system, the cortex in the brain is the answer to the problems which haunt philosophy. And when thus seen they will be seen to be in, not as marbles are in a box but as events are in a history, in a moving, growing never finished process. (Dewey 1991, p. 224)

With this focus on the context and the ongoing interaction of the organism and its surroundings, one can avoid assumptions of ontological separations. Going one step further and elaborating on the moral dimensions that Dewey expresses, neo-pragmatist Robert Brandom, in his account of intentionality, explicates the very idea of pragmatism in a way that links it to the enactivist approach to cognition: “[a] founding idea of pragmatism is that the most fundamental kind of intentionality (in the sense of directedness towards objects) is the practical involvement with objects exhibited by a sentient creature dealing skillfully with its world” (Brandom 2008, p. 178). This skillful engagement with the world is crucial for challenging prevailing paradigms surrounding correspondence theories.

The respective holistic approach envisioned by Dewey that he powerfully elaborates with his conception of *continuity* (Dewey 1934), and which is furthered by some neo-pragmatists, is reinforced by research in the neurosciences that questions the understanding of cognition as a centralized mirroring process that uses perceptual input to generate the appropriate behavioral output. Brains are studied and described as embodied, situated, and embedded.⁴

⁴ For reasons of space, I cannot discuss the rich debate around the concepts of embodiment, embeddedness, and enactivism let alone their relation to the extended mind hypothesis (for parts of the discussion see: Adams & Aizawa 2008; Clark 1997, 2001; Clark & Chalmers 1998; Rupert 2009; Shapiro 2011; Sprevak 2009; Thompson 2010; Varela et al. 1991; Ward & Stapleton 2012; Wheeler 2011). These approaches vastly differ regarding their views on representations and their general approach to cognition and action. However, each of them can offer a way of moving beyond the traditional mind-

3 Challenging the “myth of the inner” from within the Neurosciences

In the following, approaches in the empirical sciences that seek to consider the dynamic, interactionist nature of cognition will be introduced in order to enrich the view of the complexities of adaptive behaviour in self-organizing systems.

Computational cognitive neuroscientist Olaf Sporns provides a state-of-the-art synthesis of the sciences of complex networks in the brain and suggests a view beyond neurocentrism. He introduces his work as follows:

To understand these systems, we require not only knowledge of elementary systems components but also knowledge of the ways in which these components interact and the emergent properties of their interactions [...]. We cannot fully understand brain function unless we approach the brain on multiple scales, by identifying the networks that bind cells into coherent populations, organize cell groups into functional brain regions, integrate regions into systems, and link brain and body in a complete organism. (Sporns 2011, pp. 1–3)

While he does not (yet) consider the further complexities that come into play when one includes the environment of the organism, his description can be seen as a relevant, though timid first step away from a purely neurocentric view. The next step will be to recognize the relevance of environmentally attuned actions, i.e. to investigate how actions can be understood, rather than as isolated from the environment, as being in constant dynamic relation with it, adapting to requirements from the environment and in turn shaping it.

There is no doubt that the developmental perspective is crucial for understanding the dynamic interplay between social and biological processes and thus the role of the environment for experiences in developing cognition. From

body dichotomy. Specifically, enactivism focuses on the precise coupling of brain, body, and environment and might therefore be particularly promising for action-oriented approaches.

early childhood onwards, the brain is shaped by constant interaction with the world. Experiences impact on brain structure and function, as demonstrated by abundant evidence on the brain's plasticity (for classical studies, see: [Buonomano & Merzenich 1998](#), [Pascual-Leone et al. 2005](#)). Susan Oyama, in her account of developmental systems theory, argues that the mind-world dichotomy inherent in descriptions that follow dualistic accounts claiming strong gaps between the biological realm and sociocultural realm cannot do justice to evolving systems. Oyama invites us to focus on change, rather than constancy. She points to the conglomerate of heterogeneous influences that allows development. A developmental system is "a heterogeneous and causally complex mix of interacting entities and influences that produces the life cycle of an organism" ([Oyama 2000](#), p. 1). This multi-scale, interaction-driven dynamics requires an approach that does justice to context-dependency, since it is a particular context that leads to the emergence of a specific phenotype. Neglecting the context would thus necessarily lead to a failure to understand the developmental system.

Complementary to this view, Tim Ingold describes how the specificities of an environment and an organism's history with it matter for its very existence:

What goes for the relations between internal parts of the whole organism also goes for the relations between the organism and its environment. Organic forms come into being and are maintained because of a perpetual interchange with their environments not in spite of it [...]. But since an 'environment' can only be recognized in relation to an organism whose environment it is—since, in other words, it is the figure that constitutes the ground—the process of formation of the organism is the process of formation of its environment [...]. Moreover, the interface between them is not one of external contact between separate and mutually exclusive domains, for enfolded within the organism itself is the entire history of its environmental conditions. ([Ingold 1990](#), p. 216).

Consequently, rather than speaking of distinct organisms, Ingold suggests that we would be better served by speaking of the "whole-organism-in-its-environment" ([Ingold 2001](#)). In a similar way, Richard Menary suggests cognitive integration as a dynamical account of how the bodily processes of an organism in its environment lead to cognition ([Menary 2007](#)), and elaborates how manipulation of the organism's specific environment, development in that environment, and the resulting transformation of cognitive capacities in this cognitive niche matter for actual cognitive processes and our explanatory models thereof ([Menary 2010](#)).

In line with such descriptions, [Andreas Engel et al. \(2013\)](#) recently noted what they saw as a "pragmatic turn" in cognitive science, a turn that leaves aside frameworks focusing on computation over mental representation to instead study cognition as being essentially action-oriented. Building on reasoning from [Clark \(1997\)](#) and [Varela et al. \(1991\)](#), Engel and colleagues focus on the relevance of action for cognition. They discuss evidence of perception as not being neutral with respect to action but rather as part of sensorimotor couplings that are always specific for the organism, given its previous learning, experiences, and expectations. This focus implies embodiment and situatedness just as the context-sensitivity of processing. The "pragmatic turn" is based on much experimental evidence from studies on sensorimotor integration and neuronal plasticity that highlight how cognition is, in a fundamental way, grounded in action.

Taken together with many more research lines in the experimental field, these approaches can further our understanding of the essential value of what beforehand was seen to be "merely" subjective, and not necessarily real. Experience and skillful engagement with the world have a relevant, even an essential role for cognition. This insight opens the way for a more encompassing view of human experience and thus enriches Jordan and Day's account with phenomenological, anthropological, and pragmatist perspectives.

4 Why a systems approach matters

While Wild System Theory primarily seems to offer new possibilities for how to study human experiences and engagement with the world, it actually does more: it helps to develop a “theory ‘of what people are’” (Jordan & Day this collection, p. 20) by shifting our understanding of the relationships between brain, mind, body, and world. These possibilities challenge dichotomies that have for a long time dominated classical philosophical views of what human beings are and how they reason and experience. John Dewey argued against a series of dichotomies that were abundant in philosophy, such as those of mind versus body, fact versus value, internal versus external, and experience versus nature by explicating the role of continuities, e.g., between mind and body, and the importance of action for experience. A better understanding of circular causalities is necessary in order for us to be able to see humans as continually changing bodily organisms that incorporate their histories of past interactions with their environments, successful adaptations, and learning processes—each shaped their particular way of being in the world.⁵ Such a systems perspective does not seek to understand the brain in isolation, but a person in his or her idiosyncratic context.

Crucially, the approaches fostered already by John Dewey, which have today been rediscovered by philosophers and neuroscientists alike, are in fine accordance with phenomenological descriptions of what it is like to experience. How those perspectives converge into a science of mind is still to be elaborated and might receive inspiration from neurophenomenology, with its call to take seriously introspective phenomenological reports (Lutz & Thompson 2003; Varela 1996). In particular, it can be worthwhile to take this view to psychiatry, as a clinical field deeply dependent on a sensitive understanding of the relation between mind, brain,

the rest of the body, and the environment. In psychiatry it becomes particularly evident that dealing with persons is not the same as dealing with brains. For example, explaining depression as a mere chemical imbalance based on a lack of serotonin (a popular statement that does not by any means hold universally, even if one follows a strong reductionist account) does not do justice to the complex causal relationships leading to the pathology. Thomas Fuchs compellingly suggests giving up the classic physical–mental dichotomy that is present in biomedical reductionism, to develop a proper understanding of the circular causality between an organism and its environment (Fuchs 2009, 2011). Fuchs explains how an ecological concept of mental illness does justice to findings about how disorders are a product of the complex interaction of subjective, neuronal, social, and environmental influences. This does not only matter for our understanding of mental illnesses, but also importantly impacts on how we approach treatments at various levels. The essential relevance of recognizing circular causalities in the brain–body–world interaction can also be seen in neurological treatment and in the psychological reactions of patients to treatments. Beliefs about the relationship between brain and mind and how they relate to one’s personality and psychological well-being might influence reactions to neurological or neurosurgical interventions. In particular, for treatment with deep brain stimulation it has been argued that a framework that is neither dualistic nor brain-centric, but which offers a perspective that recognizes the manifold interaction between mind, body, and world can have beneficial effects on patients and their surrounding (Mecacci & Haselager 2014; Keyser & Nagel 2014). Thus, the quality of therapeutic approaches might benefit from examining more holistic approaches to psychiatric disorders and therapies. Ultimately, these theoretical considerations can be crucially relevant for life in all its facets.

5 Outlook

Abundant experimental evidence demonstrates that human beings are systems comprised of the

⁵ *The implied essentialisation of biology as a constant of human being, and of culture as its variable and interactive complement, is not just clumsily imprecise. It is the single major stumbling block that up to now has prevented us from moving towards an understanding of our human selves, and of our place in the living world, that does not endlessly recycle the polarities, paradoxes and prejudices of western thought* (Ingold 2004, p. 217).

brain as part of a body and the environment in a constant regulatory, adaptive process. Consequently, we suggest a systems view that considers such complex feedback loops in terms of circular causality (Crafa & Nagel forthcoming). As there are manifold fluctuating organismic levels that create feedback loops for continuous adaptation, studying those feedback loops will in all likelihood improve our understanding of how our experience is action-oriented and based on skillful engagement with the world. Notably, this approach does not in itself forestall by definition the assumption of representations (see e.g., Dennett 2000). I suggest that a computational view of cognition might not be opposed to the dynamic, embodied view. It is likely that we need both approaches in order to understand how self-organizing dynamic systems constantly adapting to their environment are able to reason, solve abstract problems, use language, etc (c.f., for another synthesizing suggestion, Grush 2004). Computational explanations of how the body and the environment interact can be useful tools here, possibly benefiting from ideas such as predictive coding or deep learning in Artificial Intelligence.⁶ Such a step includes blurring the boundaries between cognitive and sensory-motor processes. So-called low-level and high-level processes cannot be understood independently, since they constantly interact and influence one another. While symbolic abstraction is necessary for reasoning, problem solving, or language, those are strongly coupled to lower-level processes, such as perception, object manipulation, or movement. Much conceptual and empirical work must be undertaken, for which a mixed methods approach considering multiple dimensions seems to be necessary and most promising. Such an approach—or better, combination of approaches—can help to integrate

multiple levels of analysis. It might combine neurobiological concepts (and these on different levels as well, reaching from molecular studies up to studying systems and interacting systems) with psychological, anthropological, and philosophical studies. For the laboratory, a systems approach would ask for frameworks that allow us to study ‘active’ subjects using a variety of methods. Mobile technologies for physiological measurements are an important step towards this goal, as are set-ups that combine different physiological measurements. This is an ambitious task, which demands technological and computational innovation and effort. And, not least, studying mental capacities can be massively enriched by combining phenomenological accounts of experience with cognitive science approaches as suggested from the field of neurophenomenology (Varela 1996).

It is likely that a more holistic view on human cognition and experience will help us focus on topics that truly matter to people and that do justice to their experience. One practical consequence of a different understanding of the relationship between mind, body, and world is its potential effect on human self-understanding, which in turn can have significant psychological effects (e.g., Vohs & Schooler 2008). As Gregory Bateson frames it: “[t]he living man is thus bound within a net of epistemological and ontological premises which—regardless of ultimate truth or falsity—become partially self-validating for him” (Bateson 1973, p. 314). Thus, theoretical considerations in the field of philosophy of mind, together with the pragmatists’ understanding of experience and neuroscientific findings on the relevance of the interdependence of the brain, the rest of the body, and the environment shall lead to thicker descriptions of the multifaceted human condition.

⁶ Predictive coding is a framework for understanding the reduction of redundancy and efficient coding in the nervous system. It is suggested that highly redundant natural signals are processed by removing the predictable components of the input, thereby transmitting only what is not predictable. Hierarchical predictive coding can explain response selectivities in networks (Clark 2001; Hohwy et al. 2008; Friston et al. 2010; Friston & Stephan 2007; Rao & Ballard 1999). Inspired by neural network processing, deep learning methods in machine learning aim to produce learning of features at multiple levels of abstraction, thus allowing learning of complex functions (e.g., Arel et al. 2010; Bengio 2009; Hinton et al. 2006).

References

- Adams, F. & Aizawa, K. (2008). *The bounds of cognition*. Malden, MA: Blackwell.
- Arel, I., Rose, D. & Karnowski, T. (2010). Deep machine learning - A new frontier in artificial intelligence research. *IEEE Computational Intelligence Magazine*, 5 (4), 13-18. [10.1109/MCI.2010.938364](https://doi.org/10.1109/MCI.2010.938364)
- Bateson, G. (1973). *Steps to an ecology of mind*. London, UK: Paladin.
- Beauchamp, M. S. & Martin, A. (2007). Grounding object concepts in perception and action: Evidence from fMRI studies of tools. *Cortex*, 43 (3), 461-468. [10.1016/S0010-9452\(08\)70470-2](https://doi.org/10.1016/S0010-9452(08)70470-2)
- Bengio, Y. (2009). Learning deep architectures for AI. *Foundations & Trends in Machine Learning*, 2 (1), 1-127. [10.1561/22000000006](https://doi.org/10.1561/22000000006)
- Brandom, R. B. (2008). *Between saying and doing: Towards an analytic pragmatism*. Oxford, UK: Oxford University Press.
- Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, 47 (1-3), 139-159. [10.1016/0004-3702\(91\)90053-M](https://doi.org/10.1016/0004-3702(91)90053-M)
- Buonomano, D. V. & Merzenich, M. M. (1998). Cortical plasticity: From synapses to maps. *Annual Review of Neuroscience*, 21, 149-186. [10.1146/annurev.neuro.21.1.149](https://doi.org/10.1146/annurev.neuro.21.1.149)
- Chiel, H. J. & Beer, R. D. (1997). The brain has a body: Adaptive behavior emerges from interactions of nervous system, body, and environment. *Trends in Neuroscience*, 20 (12), 553-557. [10.1016/S0166-2236\(97\)01149-1](https://doi.org/10.1016/S0166-2236(97)01149-1)
- Clark, A. (1997). *Being there: Putting brain, body, and world together again*. London, UK: MIT Press.
- (2001). *Mindware: An introduction to the philosophy of cognitive science*. Oxford, UK: Oxford University Press.
- Clark, A. & Chalmers, D. (1998). The extended mind. *Analysis*, 58 (1), 7-19. [10.1093/analys/58.1.7](https://doi.org/10.1093/analys/58.1.7)
- Crafa, D. & Nagel, S. K. (forthcoming). Traces of culture: The feedback loop between brain, behavior, and disorder. *Transcultural Psychiatry*.
- De Jaegher, H. & Di Paolo, E. A. (2007). Participatory sense-making: An enactive approach to social cognition. *Phenomenology and the Cognitive Sciences*, 6 (4), 485-507.
- Dennett, D. C. (2000). Making tools for thinking. In D. Sperber (Ed.) *Metarepresentations: A multidisciplinary perspective* (pp. 17-29). Oxford, UK: Oxford University Press.
- Dewey, J. (1934). *Art as experience*. New York, NY: Penguin-Putnam Inc..
- (1991). Experience and nature. In J. A. Boydston (Ed.) *The later works of John Dewey (Vol. 1)*. Edwardsville, IL: Southern Illinois University Press.
- Dreyfus, H. (1992). *What computers still can't do: A critique of artificial reason*. Cambridge, MA: MIT Press.
- Engel, A. K., Fries, P. & Singer, W. (2001). Dynamic predictions: Oscillations and synchrony in top-down processing. *Nature Reviews Neuroscience*, 2 (10), 704-716. [10.1038/35094565](https://doi.org/10.1038/35094565)
- Engel, A. K., Maye, A., Kurthen, M. & König, P. (2013). Where's the action? The pragmatic turn in cognitive science. *Trends in Cognitive Science*, 17 (5), 202-209.
- Freeman, W. J. (1995). *Societies of brains: A study in the neuroscience of love and hate*. Hillsdale, NJ: Erlbaum.
- Friston, K. J., Daunizeau, J., Kilner, J. & Kiebel, S. J. (2010). Action and behavior: A free-energy formulation. *Biological Cybernetics*, 102 (3), 227-260. [10.1007/s00422-010-0364-z](https://doi.org/10.1007/s00422-010-0364-z)
- Friston, K. J. & Stephan, K. E. (2007). Free energy and the brain. *Synthese*, 159, 417-458. [10.1007/s11229-007-9237-y](https://doi.org/10.1007/s11229-007-9237-y)
- Fuchs, T. (2009). Embodied cognitive neuroscience and its consequences for psychiatry. *Poiesis and Praxis*, 6 (3-4), 219-233. [10.1007/s10202-008-0068-9](https://doi.org/10.1007/s10202-008-0068-9)
- (2011). The brain – a mediating organ. *Journal of Consciousness Studies*, 18 (7-8), 196-221.
- (2012). Are mental illnesses diseases of the brain? In S. Choudhury & J. Slaby (Eds.) *Critical neuroscience: A handbook of the social and cultural contexts of neuroscience* (pp. 331-344). Malden, MA: Wiley: Blackwell.
- Gallagher, S. (2005). *How the body shapes the mind*. Oxford, UK: Oxford University Press.
- Graziano, M. S. A., Yap, G. S. & Charles, G. G. (1994). Coding of visual space by premotor neurons. *Science*, 266, 1054-1057. [10.1126/science.7973661](https://doi.org/10.1126/science.7973661)
- Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences*, 27 (3), 377-442. [10.1017/S0140525X04000093](https://doi.org/10.1017/S0140525X04000093)
- Haken, H. (2004). *Synergetics, introduction and advanced topics*. Berlin, GER: Springer.
- Hinton, G. E., Osindero, S. & Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18 (7), 1527-1554. [10.1162/neco.2006.18.7.1527](https://doi.org/10.1162/neco.2006.18.7.1527)
- Hohwy, J., Roepstorff, A. & Friston, K. J. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, 108 (3), 687-701. [10.1016/j.cognition.2008.05.010](https://doi.org/10.1016/j.cognition.2008.05.010)

- Ingold, T. (1990). An anthropologist looks at biology. *Man, New Series*, 25 (2), 208-229.
- (1993). Tool-use, sociality and intelligence. In K. R. Gibson & T. Ingold (Eds.) *Tools, language and cognition in human evolution* (pp. 429-445). Cambridge, UK: Cambridge University Press.
- (2001). From the transmission of representations to the education of attention. In H. Whitehouse (Ed.) *The debated mind: Evolutionary psychology versus ethnography* (pp. 113-153). Berg: Oxford, UK.
- (2004). Beyond biology and culture: The meaning of evolution in a relational world. *Social Anthropology*, 12 (2), 209-221. [10.1111/j.1469-8676.2004.tb00102.x](https://doi.org/10.1111/j.1469-8676.2004.tb00102.x)
- Johnson, M. (2007). *The meaning of the body: Aesthetics of human understanding*. Chicago, IL: University of Chicago Press.
- Jordan, J. S. & Day, B. (2015). Wild systems theory as a 21st century coherence framework for cognitive science. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1-21). Frankfurt a. M., GER: MIND Group.
- Kelso, J. A. S. (1995). *Dynamic patterns*. Cambridge, MA: MIT Press.
- Keyser, J. & Nagel, S. K. (2014). Stimulating more than the patient's brain: Deep brain stimulation from a systems perspective. *American Journal of Bioethics*, 5 (4), 60-62. [10.1080/21507740.2014.953268](https://doi.org/10.1080/21507740.2014.953268)
- Kitcher, P. S. (2012). *Preludes to pragmatism: Toward a reconstruction of philosophy*. Oxford, UK: Oxford University Press.
- Lakoff, G. & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to western thought*. New York, NY: Basic Books.
- Lewis, M. D. (2005). Bridging emotion theory and neurobiology through dynamic systems modeling. *Behavioral and Brain Sciences*, 28 (2), 169-245. [10.1017/S0140525X0500004X](https://doi.org/10.1017/S0140525X0500004X)
- Lutz, A. & Thompson, E. (2003). Neurophenomenology: Integrating subjective experience and brain dynamics in the neuroscience of consciousness. *Journal of Consciousness Studies*, 10 (9-10), 31-52.
- Maturana, H. R. & Varela, F. J. (1980). *Autopoiesis and cognition: The realization of the living*. Dordrecht, NL: Reidel.
- Mecacci, G. & Haselager, W. F. G. (2014). Stimulating the self: The influence of conceptual frameworks on reactions to deep brain stimulation. *AJOB Neuroscience*, 5 (4), 30-39. [10.1080/21507740.2014.951776](https://doi.org/10.1080/21507740.2014.951776)
- Menary, R. (2007). *Cognitive integration: Mind and cognition unbounded*. Basingstoke, UK: Palgrave Macmillan.
- (2010). Dimensions of mind. *Phenomenology and the Cognitive Sciences*, 9, 561-578.
- Merleau-Ponty, M. (1962). *Phenomenology of perception*. London, UK: Routledge & Kegan Paul.
- Noë, A. (2004). *Action in perception*. Cambridge, UK: MIT Press.
- O'Regan, J. K. & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24 (5), 939-1031. [10.1017/S0140525X01000115](https://doi.org/10.1017/S0140525X01000115)
- Oyama, S. (2000). *Evolution's eye: A systems view of the biology-culture divide*. Durham, NC: Duke University Press.
- Pascual-Leone, A., Amedi, A., Fregni, F. & Merabet, L. B. (2005). The plastic human brain cortex. *Annual Review of Neuroscience*, 28, 377-401. [10.1146/annurev.neuro.27.070203.144216](https://doi.org/10.1146/annurev.neuro.27.070203.144216)
- Rao, R. P. N. & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2, 79-87. [10.1038/4580](https://doi.org/10.1038/4580)
- Rupert, R. (2009). *Cognitive systems and the extended mind*. New York, NY: Oxford University Press.
- Schulkin, J. (2009). *Cognitive adaptation: A pragmatist perspective*. Cambridge, UK: Cambridge University Press.
- Shapiro, L. (2011). *Embodied Cognition*. New York, NY: Routledge.
- Sporns, O. (2011). *Networks of the brain*. Cambridge, MA: MIT Press.
- Sprevak, M. (2009). Extended cognition and functionalism. *The Journal of Philosophy*, 106 (9), 503-527.
- Thelen, E. & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action*. Cambridge, MA: MIT Press.
- Thompson, E. (2010). *Mind in life: Biology, phenomenology, and the sciences of mind*. Cambridge, MA: MIT Press.
- Torrance, S. (2009). Contesting the concept of consciousness. *Journal of Consciousness Studies*, 16 (5), 111-126.
- Varela, F. J. (1996). Neurophenomenology: A methodological remedy for the hard problem. *Journal of Consciousness Studies*, 3 (4), 330-350.
- Varela, F. J., Thompson, E. & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, MA: MIT Press.
- Vohs, K. D. & Schooler, J. W. (2008). The value of believing in free will: Encouraging a belief in determinism increases cheating. *Psychological Science*, 19 (1), 49-54. [10.1111/j.1467-9280.2008.02045.x](https://doi.org/10.1111/j.1467-9280.2008.02045.x)

- von Uexküll, J. (1940). Bedeutungslehre. *Semiotica*, 42 (1), 25-82.
- Ward, D. & Stapleton, M. (2012). Es are good: Cognition as enacted, embodied, embedded, affective and extended. In F. Paglieri (Ed.) *Consciousness in interaction: The role of the natural and social context in shaping consciousness* (pp. 89-104). Amsterdam, NL: John Benjamins.
- Wheeler, M. (2011). Embodied cognition and the extended mind. In J. Garvey (Ed.) *The continuum companion to philosophy of mind* (pp. 220-238). London, UK: Continuum.
- Wilson, M. & Knoblich, G. (2005). The case for motor involvement in perceiving conspecifics. *Psychological Bulletin*, 131 (3), 460-473. [10.1037/0033-2909.131.3.460](https://doi.org/10.1037/0033-2909.131.3.460)

After Naturalism: Wild Systems Theory and the Turn To Holism

A Reply to Saskia K. Nagel

J. Scott Jordan & Brian Day

We agree with Dr. Nagel's assertion that explanations within cognitive science can be *thickened* by an infusion of pragmatism and anthropology. We further propose that because of its direct challenge of the correspondence thinking that tends to underlie contemporary indirect- and direct realism, Wild Systems Theory provides a *coherence* framework that conceptualizes reality as inherently context dependent and, therefore, inherently *meaning-full*. As a result, pragmatists can appeal to the reality of lived experience, anthropologists can appeal to the meaningful, multi-scale influences that shape an individual, and both can do so without having to justify the reality status of meaning in relation to the meaning-less view of reality we have been led to via the indirect- and direct-realism inherent in contemporary naturalism.

Keywords

Coherence theory of truth | Correspondence theory of truth | Direct realism | Embodiment | Epistemic gap | Indirect realism | Intrinsic properties | Modes of experience | Multi-scale self-sustaining systems | Reality | Wild systems theory

"We are caught up in an inescapable network of mutuality..."
Dr. Martin Luther King, Jr., 1964

1 Introduction

In her commentary on our paper, Dr. Saskia Nagel calls for a thickening of the descriptions we give in cognitive science. By *thickening* she means, ...a dense description specifying details and patterns and considering contextual factors, of human experience and cognition. ([Nagel this](#)

[collection](#), p. 3). Dr. Nagel further asserts that one way to achieve such a thickening is to infuse cognitive science with the views of pragmatism (i.e., John Dewey) and anthropology (i.e., Timothy Ingold). We couldn't agree more, and we applaud Dr. Nagel's appeal to Dewey and

Authors

[J. Scott Jordan](#)

jsjorda@ilstu.edu

Illinois State University

Bloomington-Normal, IL, U.S.A.

[Brian Day](#)

bmday15@gmail.com

Clemson University

Clemson, SC, U.S.A.

Commentator

[Saskia K. Nagel](#)

s.k.nagel@utwente.nl

University of Twente

Enschede, Netherlands

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität

Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University

Melbourne, Australia

Ingold as a means of allowing multi-scale contextual factors to play a much larger role in our accounts of cognition and consciousness.

Given our agreement on the important contributions that pragmatism and anthropology can make to cognitive science, we also feel the need to express our belief that WST (Wild Systems Theory) and its conceptualization of organisms as *self-sustaining embodiments of context* (versus physical-mental, or mind-body systems) actually creates a conceptual framework within which the views of Dewey and Ingold can move beyond the conceptual constraints of contemporary pragmatism and anthropology.

2 Pragmatism and Wild Systems Theory

In a recent paper regarding WST, [Jordan & Vinson \(2012\)](#) propose that Dewey's brand of pragmatism represented a rather unique combination of an idealist approach to metaphysics and an epistemic (i.e., pragmatic) approach to science. Specifically, Dewey's early training as an idealist philosopher led him to reject the objective-subjective, correspondence-driven approach to reality and truth that was prominent in the *indirect*- and *direct-realist* versions of naturalism that were emerging during his time. Instead, Dewey believed, as did his idealist, *coherentist* mentors, that meaning and value were *constitutive* of reality. In addition, given his *coherence*- (versus *correspondence*-) driven metaphysics, Dewey believed that science was a practice that afforded us the opportunity to reveal patterns of contingency within the contexts in which we are embedded. He repeatedly emphasized this epistemic, pragmatic approach to science as a way to challenge the more ontologically minded, metaphysical approach to science that was being espoused by indirect- and direct-realist forms of naturalism:

The search for 'efficient causes' instead of for final causes, for extrinsic relations instead of intrinsic forms, constitutes the aim of science. But the search does not signify a quest for reality in contrast with experience of the unreal and phenomenal.

It signifies a search for those relations upon which the *occurrence* of real qualities and values depends, by means of which we can regulate their occurrence. To call existences as they are directly and qualitatively experienced 'phenomena' is not to assign to them a metaphysical status. It is to indicate that they set the problem of ascertaining the relations of interaction upon which their occurrence depends. ([Dewey 1929](#), pp. 103-104)

Despite Dewey's concerns, his unique combination of idealist ontology and scientific pragmatism eventually gave way to what [Gardner \(2007\)](#) refers to as the *Hard Naturalism* of our time, in which meaning and value are seen as completely unnecessary in a scientific, causal description of reality:

By the time we get to Freud ... let alone Quine, naturalism is conceived as resting exclusively on theoretical reason and as immune to non-theoretical attack—it is assumed that nothing could be shown regarding the axiological implications of naturalism that would give us reason to reconsider our commitment to it: we have ceased to think that naturalism is essential for the realization of our interest in value, and do not believe that it would be an option for us to reject naturalism even if it were to prove thoroughly inimical to our value-interests. (p. 24)

Within the contemporary context of Hard Naturalism, pragmatic philosophers such as [Richard Shusterman \(2008\)](#) tend to downplay and even eschew ontology. Specifically, Shusterman asserts that 20th century ontological approaches to the mind and body that were espoused by the likes of William James and Merleau-Ponty actually led us to devalue bodily sensations in the name of developing our rational capacities.

Merleau-Ponty's commitment to a fixed, universal phenomenological ontology based on primordial perception thus provides further reason for dismissing the value of

explicit somatic consciousness. Being more concerned with individual differences and contingencies, with future-looking change and reconstruction, with pluralities of practice that can be used by individuals and groups for improving on primary experience, pragmatism is more receptive to reflective somatic consciousness and its disciplinary uses for philosophy. (Shusterman 2008, p. 66)

Clearly, there are important continuities between the pragmatic philosophies of Dewey and Shusterman (Jordan 2010). Specifically, Shusterman's focus on *practice* overlaps with Dewey's conceptualization of science as a practice as opposed to a tool for metaphysics. In addition, Shusterman's emphasis on *primary experience* is consistent with Dewey's idealist commitment to the reality of experience. The major difference between the two seems to be Shusterman's lack of interest in, or perhaps outright disdain for metaphysics.

One possible reason for Shusterman's (2008) lack of interest in metaphysics may be our contemporary commitment to Hard Naturalism. As was stated in the quotation by Gardner (2007), Hard Naturalism seems so implicitly accepted these days, it seems difficult, if even possible, to propose a metaphysics in which value, meaning, and experience are constitutive of reality. Because of its commitment to the reality of experience however, as well as its clear questioning of the indirect- and direct-realism that lie at the core of Hard Naturalism, WST seems perfectly situated to take-up Dewey's anti-correspondence arguments and place them within a 21st century coherentist framework. Instead of remaining within the centuries-old conceptual framework of *mind* and *body* however, as Dewey did, WST takes the philosophical risk of creating a new concept: specifically, *embodied context*. We say *philosophical risk* because the notion of embodied context conceptualizes meaning in the exact opposite fashion as Hard Naturalism. Specifically, it renders meaning ubiquitous throughout reality. Given the century of philosophical work that has ultimately led to the Hard Naturalist belief that reality is inher-

ently meaningless, we suspect some might see it as simply silly or heretical to assert that reality is inherently meaningful, through and through. This is why we consider the concept of *embodied context* risky. Regardless of the risks however, we see WST as a means of getting meaning back into reality. It does so by following the lead of the idealists, particularly Oakeshott (1933), who did not appeal to the a priori, the transcendental, or the absolute, and refused to describe reality in terms of the observer-independent intrinsic properties that ultimately make it difficult, if not logically impossible, for meaning to be constitutive of reality. Within WST's coherentist perspective, Dewey's pragmatism is restored as a 21st century framework, and pragmatism, in general, can commit itself to the reality of lived experience in an ontological fashion that does not require justification in relation to Hard Naturalism.

To be sure, there have been those scholars who have attempted to introduce meaning back into Hard Naturalism by referring to it via terms such as *emergent* and *irreducible*. Gardner (2007) however, refers to such attempts as *Soft Naturalism* and states the following:

If, then, it is demonstrated successfully by the soft naturalist that such-and-such a phenomenon is not reducible to the natural facts austere conceived, this conclusion is not an end of enquiry, but rather a reaffirmation of an explanandum, i.e., a restatement that the phenomenon stands in need of metaphysical explanation. Irreducibility arguments, if successful, yield data that do not interpret or explain themselves, but call for interpretation: the soft naturalist needs to say something on the subject of why there should be, in general, phenomena that have substantial reality, but do not owe it to the hard natural facts. (p. 30)

WST avoids collapsing into Soft Naturalism because it directly challenges the Hard Naturalist assumption of intrinsic, context-independent properties. It does so by asserting that all properties are necessarily context-dependent and

thus, inherently meaning-full. In short, meaning is constitutive of reality.

3 Anthropology and Wild Systems Theory

In addition to providing a contemporary framework for pragmatism, WST also provides a straightforward means of integrating cognitive science and anthropology. For example, in her comment on our paper Dr. Nagel points to the work of [Timothy Ingold](#) as a contemporary example of an anthropologist whose work can *thicken* our understanding of cognition and experience.

Knowing does not lie in the establishment of a correspondence between the world and its representation, but is rather immanent in the life and consciousness of the knower as it unfolds within the field of practice set up through his or her presence as a being-in-the-world. (2011, p. 159)

While WST couldn't agree more with [Ingold's](#) (2011) critique of correspondence approaches to the nature of knowledge, WST's conceptualization of living systems as multi-scale, self-sustaining embodiments of the phylogenetic, cultural, social, and ontogenetic contexts within which they emerged and within which they sustain themselves provides a straight forward explanation of why *knowing* is, "...immanent in the life and consciousness of the knower..." ([Ingold 2011](#), p. 159). Specifically, knowing is immanent in *being-in-the-world* because organisms, as embodiments of context, *are* knowledge ([Jordan 2000](#)). In short, they are *world in world*. Thus, as implied by Ingold, to *be* is to *mean*.

A potential advantage of WST's approach to this issue is that it directly addresses the Hard Naturalism that underlies the correspondence-driven thinking [Ingold](#) (2011) critiques. That is, by problematizing the realist assumption of context-independent, intrinsic properties, WST asserts it is logically impossible for meaningless *things* to exist. That is, it is logically impossible to *be* and *not mean*. By engaging in this ontological spadework, WST does not suffer

the risk of collapsing into Soft Naturalism, as does Ingold's position, or any position for that matter, that attempts to establish the reality of experience without addressing Hard Naturalism's assertion that meaning is not constitutive of reality.

In addition to addressing [Ingold's](#) (2011) *being-in-the-world* approach to meaning, WST also addresses Dr. Nagel's assertion that anthropology can *thicken* cognitive science by leading us to consider the continuous, un-ending influence that multiple scales of context (e.g., phylogenetic, cultural, social, and ontogenetic) have on the nature of bodies and meaning. She develops this point by referring to [Susan Oyama's](#) (1985) assertion that in addition to inheriting genes, infants also inherit a heterogeneous collection of multi-scale contexts, including other persons, that continuously shape, and are shaped by, the developing individual. Oyama refers to this collection of contexts as a *developmental system*. While describing Oyama's work, Dr. [Nagel](#) states:

This multi-scale, interaction-driven dynamics requires an approach that does justice to context-dependency, since it is a particular context that leads to the emergence of a specific phenotype. Neglecting the context would thus necessarily lead to a failure to understand the developmental system. ([this collection](#), p. 6)

Again, we couldn't agree more with Drs. Nagel and Oyama. What WST potentially adds to the notion of a developmental system is the idea that self-sustaining systems constitute embodiments of their developmental contexts. The advantage here is the same advantage we encountered when addressing WST's relationship to [Ingold's](#) (2011) *being-in-the-world* approach to meaning. By providing a coherentist ontology that renders reality inherently meaningful, WST constitutes a meaningful alternative to Hard Naturalism's correspondence-driven assertion that reality is inherently meaningless. As a result, WST allows one to utilize [Oyama's](#) (1985) notion of *developmental contexts* in a way that prevents one from having to explain how it is that developmental contexts

render an inherently meaningless reality meaningful. Specifically, developmental contexts don't have to render meaningless reality meaningful because, according to WST, all phenomena are context dependent and, therefore, inherently meaningful.

4 Conclusions

In the end, we agree with Dr. Nagel's assertion that pragmatism and anthropology provide a means of *thickening* our descriptions of bodies and meaning. We further propose that WST helps achieve such a *thickening* because it asserts that bodies (i.e., embodied contexts) *are* meaning. From this perspective, anthropology and cognitive science both involve the study of meaning, and differ only in that they focus their descriptions on different levels of nested context, or, to say it another way, different levels of nested meaning.

In addition to providing a means of integrating cognitive science and anthropology, WST's focus on a coherence approach to truth, as opposed to a correspondence approach to truth, puts it in a position to provide an integrative framework for scholarship in general (Jordan & Vandervert 1999; Jordan & Vinson 2012). In short, all disciplines study some scale of reality, and any scale being measured, because of its inescapable context dependence, is inherently meaningful. This observation leads to yet another point at which we are in agreement with Dr. Nagel. Specifically, we very much appreciate her assertion that WST helps to develop a different approach to *what people are*. By modeling all of reality as context-dependent, and self-sustaining systems as embodiments of context, WST conceptualizes each and every one of us as *world in world* instead of as meaningless physical systems. As a result, we are all inescapably meaningful and efficacious. Everything we do alters the contexts within which we sustain ourselves. Everything we do matters.

Given WST's ability to provide a means of bypassing the meaningless view of reality we have been led to via Hard Naturalism, it is not clear to what extent philosophy is so much ex-

periencing a *pragmatic* turn (Engel et al. 2013) as it is experiencing a *holist* turn (Jordan 2013). If it proves to be the latter, sustaining such a turn will be difficult, for it will force us to experience our scientific concepts (e.g., physical, chemical, biological) as epistemic tools we must necessarily utilize if we are to get on with the cooperative, social practice of science. As was stated by Oakeshott (1933) however, science as a mode of experience is inherently an abstraction, an arrestment from the whole. This means that while the practice of science necessitates that we generate conceptual abstractions regarding that within which we are nested, we must always remember that our abstractions can never satisfy a correspondence-driven definition of truth. In short, while we must necessarily represent, we must simultaneously commit to uncertainty. Perhaps it was the potential pathos of this conundrum that W. G. Sebald was referring to in his poem *After Nature*:

For it is hard to discover
the winged vertebrates of prehistory
embedded in tablets of slate.
But if I see before me
the nervature of past life
in one image, I always think
that this has something to do
with truth. Our brains, after all,
are always at work on some quivers
of self-organization, however faint,
and it is from this that an order
arises, in places beautiful
and comforting, though more cruel, too,
than the previous state of ignorance
(2003, p. 2)

References

- Dewey, J. (1929). *The quest for certainty: The study of the relation of knowledge and action*. New York, NY: Minton, Balch & Company.
- Engel, A. K., Maye, A., Kurthen, M. & König, P. (2013). Where's the action? The pragmatic turn in cognitive science. *Trends in Cognitive Sciences*, 17 (5), 202-209. [10.1016/j.tics.2013.03.006](https://doi.org/10.1016/j.tics.2013.03.006)
- Gardner, S. (2007). The limits of naturalism and the metaphysics of German idealism. In E. Hammer (Ed.) *German idealism: Contemporary perspectives* (pp. 19-49). Abingdon, UK: Routledge.
- Ingold, T. (2011). *Being alive: Essays on movement, knowledge and description*. New York, NY: Routledge.
- Jordan, J. S. (2000). The world in the organism: Living systems are knowledge. *Psychology*, 11 (113)
- (2010). Shusterman, Merleau-Ponty, and Dewey: The role of pragmatism in the conversation of embodiment. *Action, Criticism, and Theory for Music Education*, 9 (1), 67-73. http://act.maydaygroup.org/articles/Jordan9_1.pdf
- (2013). Consciousness and embodiment. In H. Pashler (Ed.) *The encyclopedia of the mind*. Thousand Oaks, CA: Sage Reference.
- Jordan, J. S. & Vandervert, L. (1999). Liberal education as a reflection of our assumptions regarding truth and consciousness: Time for an integrative philosophy. In J. S. Jordan (Ed.) *Modeling consciousness across the disciplines* (pp. 307-331). New York, NY: University Press of America.
- Jordan, J. S. & Vinson, D. (2012). After nature: On bodies, consciousness, and causality. *Journal of Consciousness Studies*, 19 (5-6), 229-250.
- Nagel, S. (2015). Thickening descriptions with views from pragmatism and anthropology-A Commentary on Scott Jordan and Brian Day. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a.M., GER: MIND Group.
- Oakeshott, M. (1933). *Experience and its modes*. Cambridge, UK: Cambridge University Press.
- Oyama, S. (1985). *The ontogeny of information: Developmental systems and evolution*. Cambridge, UK: Cambridge University Press.
- Sebold, W. G. (2003). *After nature*. New York, NY: The Modern Library.
- Shusterman, R. (2008). *Body consciousness: A philosophy of mindfulness and somaesthetics*. Cambridge, UK: Cambridge University Press.

The Crack of Dawn

Perceptual Functions and Neural Mechanisms that Mark the Transition from Unconscious Processing to Conscious Vision

Victor Lamme

There is conscious vision, and there is unconscious visual processing. So far so good. But where lies the boundary between the two? What are the visual functions that shape the transition from “processing in the dark” to having a conscious visual percept? And what are the neural mechanisms that carry that transition? I review the findings on feature detection, object categorization, interference, inference, Gestalt grouping, and perceptual organization, and examine to what extent these functions correlate with the presence or absence of conscious vision. It turns out that a surprisingly large set of visual functions is executed unconsciously, indicating that unconscious vision is much “smarter” than we might intuitively think. Only when these unconscious mechanisms fail, and more elaborate and incremental processing steps are required, is consciousness necessary. The function of conscious vision may be to add a final layer to our interpretation of the world, to solve relatively “new” visual problems, and to enable visual learning.

Keywords

Access | Anaesthesia | Attention | Consciousness | Continuous flash suppression | Feature detection | GABA | Gestalt laws | Human | Masking | Monkey | NMDA | Object categorization | P-consciousness | Perceptual inference | Perceptual interference | Perceptual organization | Phenomenal experience | Qualia | Report | Rivalry | The hard problem | Visual cortex | Visual perception

Author

[Victor Lamme](#)

Victorlamme@gmail.com

Universiteit van Amsterdam
Amsterdam, Netherlands

Commentator

[Lucia Melloni](#)

lucia.melloni@brain.mpg.de

Max Planck Institute for Brain
Research
Frankfurt a. M., Germany

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Qualia 2.0

What do we need to know about consciousness? Which aspect of it is most mysterious? What do we want philosophy, psychology, neuroscience, computer science, or even physics to tell us about consciousness that we do not already know? The answer to that question may vary from person to person. To me it is this very simple thing: why do I see? Why do I have conscious experiences whenever I open my eyes? What makes the 1.5 kilograms of protein and fat in my head give me the wonderful sensations I experience every day,

from the second I wake up until the moment I fall asleep?

The point is probably best illustrated by the difference between a digital photo camera and the human mind. A camera nowadays can do wonderful things. It can record an image at extreme resolutions, with the right focus and exposure, all by itself. It can identify a face, putting it in a frame on the screen, and writing the name below it of the person it recognizes. You can leave it to push the button at the moment everybody smiles. Connect it to a com-

puter, and it will detect emotions, recognize objects, or read handwriting on a letter. Surveillance cameras can detect suspicious movements or strange behaviours in crowds, outperforming human night-guards or intelligence agents. There is one big difference between the camera and the human mind, though. The camera does not see.¹ I do. And so does the night guard, most of the time. It is this aspect of visual processing that is in need of an explanation. Not the fact that I recognize the person in front of me, can read his emotions, talk to him, or pick up the cup of coffee he gives me. I can vaguely understand how my brain enables me to do that. What I do not understand is how it is that I see all those things.

Is that the “hard problem” all over? Am I talking Qualia? Not in the strict sense. In its original formulation, the hard problem would argue that there is no function, no neural process whatsoever that could ever explain conscious sensations (Chalmers 1995). Functions explain functions, but not the fact that I see. Qualia are defined as ineffable aspects of information: the redness of red, stripped of every possible functional property or reactive disposition. And with that comes the whole charade of inverted spectra, colour scientists called Mary, and explanatory gaps. Which didn’t get us all that far—so let’s not chase that unicorn again.

It’s not that I don’t want to address the hard problem, or bridge the explanatory gap. That is in fact exactly what I am after (Lamme 2010a, 2010b). But I would like to leave that for later. What we need to recognize, first, is that there must be some functions and some neural processes that are more closely connected to seeing than others (Crick & Koch 1998, 2003). For example, it is fairly reasonable to assume that an understanding of the neural basis of a reflexive motor response—like the pulling away of your hand when it touches fire—does very little towards explaining consciousness (Lamme & Roelfsema 2000; Lamme 2006). Other func-

tions may offer a better gateway. For example because they explain some fundamental aspect of seeing (Seth 2010), such as its unity, or because they coincide with the difference between conscious visual processing and visual processing that occurs “in the dark” (Lamme 2010a, 2010b). In trying to bridge the explanatory gap, I think it is important to first find the right tree up which to bark. We must first identify the exact boundaries between conscious and unconscious processing. The hard problem can then be attacked afterwards. Or maybe that whole explanatory gap will vanish right before our eyes once we are there.

This paper is about exactly that. Let’s find the visual functions and neural processes that take us as close as possible to the hard problem, as close as possible towards explaining why we humans see, while photo cameras do not. And let’s avoid barking up the wrong tree.

2 Why dolphins are not fish

To find the cognitive functions and neural processes that take us towards understanding the phenomenality of consciousness it is important to establish a boundary—a boundary with what we should call unconscious processing. This will by no means be an easy job (Lamme 2006). In fact, the whole issue of understanding consciousness and solving the explanatory gap is about positioning that boundary. There are situations where it is in fact unclear whether we should talk about a conscious sensation or not. Take the situation of a split-brain patient: here, a stimulus presented to the left visual field will be processed by the right half brain, typically devoid of communication via language. Hence, the subject will *tell* you that she did not see that stimulus. He may draw the stimulus, however, using his left hand. Or the left hand may point at the stimulus, or match it to a related subject (Gazzaniga 2005). Who are we to believe, then? The hand or the mouth? What types of behaviour may count as evidence for conscious sensations? Just speech? What about aphasic subjects, then? Similarly, there are conditions like neglect, or manipulations of attention (change blindness, inattention blindness),

¹ Or at least we assume it does not. This is the basic intuition we start from in trying to explain consciousness. If not, one easily slides into pan-psychism. That is a viable option of course: it could be that the camera does see, yet cannot “tell” us. However, the arguments put forward in the remainder of this paper seem to suggest that the camera does not see.

where it is difficult to be entirely sure that what appears to be not seen is in fact maybe just not attended to, and hence forgotten or not cognitively accessible and hence not reportable (Lamme 2003, 2006, 2010). This uncertainty has sparked a lively debate on the nature of consciousness, its potential independence of cognitive functions like attention, working memory, or access (Lamme 2004, 2010a, 2010b), and whether consciousness can ever be separated from a *report* about consciousness (Block 2005, 2007; Dehaene et al. 2006; Cohen & Dennett 2011; Fahrenfort & Lamme 2012). This debate is all about the difference between seeing and knowing, between phenomenality and access (P-consciousness and A-consciousness), between qualia and higher-order thoughts. In this debate, the issue that seems unsolvable is where exactly the boundary between conscious and unconscious processing should be laid.

In such attempts to establish boundaries, it is perhaps good to start from the extremes, as an example from zoology will illustrate. Superficially, one could argue about whether a dolphin is a fish or a mammal. Science has resolved that argument by looking at animals that we all agree are either mammals (such as dogs, cows, or monkeys) or fish (such as sole, tuna, or piranha). From that perspective, the key differences between these species lies in the way they breathe and reproduce.² Why are these the key differences? Well, differences in breathing do all the explaining for why fish are generally more adept at living in water instead of on land. Similarly, evolution towards the land has called for eggs with protective layers (amnios), as anamniotic eggs (that fish lay) cannot survive on land. The most extreme version of that is the intrauterine development of the egg. Mammals and fish are thus at the two extreme ends of evolutionary adaptation towards breathing and reproducing on land.³ We understand why a mammal behaves differently to a fish from these key properties. From these key properties we understand why mammals roam the surface of the earth, why they look the way they look, and

why they behave the way they do.⁴ In classifying animals, we use these features for a discrete taxonomy. This means that there are other features that do not qualify as defining characteristics, which are disregarded in animal taxonomy. Among these are behaviours like swimming in water, or living in groups. The key differences, obtained from looking at the extreme ends of the spectra, lead us to conclude that dolphins are mammals and not fish, even though appearances may suggest otherwise. We can draw a sharp boundary, and do not have to resort to saying that dolphins are “fishy mammals”, because we recognize that the swimming behaviour of dolphins is irrelevant to their taxonomy.⁵ *Defining features* and *irrelevant features* enables a proper and discrete taxonomy, making most sense of all the available data. Moreover, a taxonomy based on such features allows for an understanding that goes towards a deeper level, in this case the evolutionary pressure that came from the transition from sea to land dwelling.

I propose to undertake something similar with consciousness. What is the proper taxonomy of conscious versus unconscious vision? What are the defining features of this difference, and what features are irrelevant? And do the defining features take us towards a somewhat more fundamental level of understanding consciousness (Lamme 2010a, 2010b)? To find those features, we start from the extreme ends: the mammals and fish of consciousness research, the things most people will agree on as representing either conscious or unconscious processing.

3 The mammals and fish of consciousness

When I am awake and say I see a face, am able to report its identity; I can identify the colour of its eyes and hair, and judge its emotional ex-

² Among other things, like whether they maintain body temperature or have hairy skin.

³ With reptiles and birds in between, laying amniotic eggs on the land.

⁴ Of course there is the occasional mammal that lays eggs (e.g., the platypus) or fish that give birth to live young (e.g., the hammerhead shark). Still, calling these mammals or fish depends on the relative weight of other defining features, such as their way of breathing, feeding, body temperature maintenance, etc.

⁵ In a somewhat more mathematical analogy one could take all properties of all animals in the world, and perform a cluster or factor analysis. A good taxonomy has clusters that are aligned along the primary factors. Traditional taxonomy seems to have operated in this way implicitly.

pression. There is little reason to doubt that I have a conscious sensation of that face.⁶ If we study the properties of visual processing in this condition, we can be pretty sure we are studying the properties of conscious visual perception. This is our “mammal” of consciousness. We can study the properties of this species fairly easily. We can resort to introspection, verbal reports, or more strictly formalized approaches like detection or discrimination tasks. In favour of using introspection is that our introspective idea of consciousness is the very thing we are trying to explain. I would like to understand why the world looks the way it looks in my mind’s eye. This is the explanandum. Even so, we should be cautious in fully “trusting” introspection,⁷ and that is where more formal approaches may come in handy.

What would be the proper “fish” of consciousness? Are there conditions where everyone agrees that consciousness is absent? Dreamless sleep (Tononi & Massimini 2008) and anaesthesia (Alkire et al. 2008) seem to be good candidates, although not very useful ones, given that visual stimuli are difficult to deliver, and that one can only resort to objective measures (brain signals) to assess what is still processed or not. Awake subjects are easier to assess in that respect, but there it is hard to find truly unequivocal manipulations of consciousness. “Unequivocal” in this context means that the manipulation can truly be regarded as a manipulation of consciousness, i.e., in the case of vision is a manipulation of visibility (Kim & Blake 2005; Lamme 2006). An example of the latter

would be visual masking (Breitmeyer & Ogmen 2000; Enns & Di Lollo 2000). Here, a target stimulus is presented very briefly, and immediately followed by another stimulus, known as the mask. When properly done, this will render the target completely invisible. People will be at chance detecting presence or absence, or in judging another property of the target stimulus. It is safe to assume invisibility in masking, because there is no conceivable reason that could prevent the subject from reporting his visual percept, had he had one: the subject is sitting there, focussing his full attention to the target location, ready to push the button as soon as he sees the target. The not-seeing can therefore not be attributed to the absence of attention, to a lapse of memory, or to any other cognitive function sitting between a potentially conscious sensation and its report (Lamme 2003, 2010a, 2010b). As we are ready to believe the presence of consciousness in the case of someone verbally describing the face he sees, we should be equally ready to believe its absence in the case of masking (or dreamless sleep and anaesthesia).⁸

Another popular paradigm to render stimuli invisible is continuous flash suppression (CFS; Tsuchiya & Koch 2005). Here, the target stimulus is shown to one eye, while the other eye receives a rapid stream of brightly coloured patches, serving as a mask. This typically results in the target stimulus being rendered invisible, although stimuli may “break through” after a while.⁹ A third paradigm is dichoptic masking, where two oppositely coloured stimuli are shown to the two eyes, that when properly fused combine into an invisible stimulus (Mout-

6 One could do so, of course, which would lead to the denouncement of consciousness as a scientific phenomenon altogether, much along the lines of eliminative materialism (e.g., Churchland 1981). Daniel Dennett, in his categorical denouncement of anything coming close to qualia or even the phenomenology of consciousness, seems to follow a similar agenda (1993). It is entirely possible indeed that consciousness is a figment of our imagination, one that will evaporate upon close scientific scrutiny. Something like that happened to ‘elan vital’—the unique property of living matter—once we learned about chemistry, biology, DNA, and natural selection. For now, let’s assume that consciousness exists, and is in need of an explanation. If not, I would rather not be spending my years in neuroscience.

7 One important caveat is that for introspection we have to resort to cognitive functions like attention, memory, and “internal report”. This may result in both a potential underestimation of what we actually see (see for example the iconic/fragile/working memory discussion), and to an overestimation of what we actually see (as in the illusion of peripheral colour vision). This has been dealt with extensively elsewhere (Lamme 2010a).

8 Note that a proper treatment of response bias is important in this case. “Shy” subjects may feel inclined to respond “not seen” on most trials, more liberal subjects may feel inclined to respond “seen” on most trials. Only treating the responses in terms of signal detection theory (Swets et al. 1978) can truly establish the absence of any sensation (because the number of false alarms—subjects saying “seen” on trials without a target—is taken into account). From that perspective, using only partially-effective masks is not a proper method, not even when only those trials are used in which subjects reported not seeing the target.

9 A potential problem with the CFS manipulation is that “time to breakthrough” is often used as a measure of relative awareness of stimuli. Time to breakthrough is more or less analogous to a “yes” response (or hit) in a masking paradigm, and hence can suffer from response bias. CFS studies where responses are more rigorously treated in terms of signal detection theory are scarce. See Stein et al. (2011) for a more elaborate discussion on this problem with the CFS paradigm.

oussis & Zeki 2002; Fahrenfort et al. 2012). From all the available neuropsychological patients, patients suffering from hemianopia due to a V1 lesion (often accompanied by blindsight) are probably the clearest cases of impaired visual consciousness (Weiskrantz 1996).

I select these consciousness manipulations because they seem to be the safest bets for highlighting situations where conscious vision is really absent. The absence of conscious vision in these cases has purely visual origins. There is no other function precluding the report of a potentially present visual sensation, as may be the case in split-brain patients or neglect, or in manipulations like inattentive blindness, change blindness, or the attentional blink (Lamme 2003). The two extreme ends—the mammals and fish of consciousness—may serve as a guideline towards establishing the properties of conscious versus unconscious processing. What are the differences between awake conscious vision and vision in sleep, anaesthesia, blindsight, and the various forms of masking?

4 Categorization: From low to high level features

Above, I used the example of seeing a face. What does seeing a face mean, in terms of the visual functions being executed? Recognizing a face first of all entails that one identifies the stimulus as belonging to the class “faces”, as opposed to any other class of stimuli, such as “animals”, “teapots”, “houses”, or “letters”. This is a process of categorization. Intuitively, categorization seems a key property of consciously seeing and recognizing a face. It is not, however. Since the first findings of blindsight it has been recognized that categorization can occur fully independently of conscious sensations (Weiskrantz 1996; Boyer et al. 2005). Patients without awareness of stimuli in the blind part of the visual field can nevertheless categorize these stimuli, as long as the categorization is framed in a two-alternative forced choice: is it a square or a circle, is it moving upwards or downwards, is it red or green, vertical or horizontal? In such cases, patients’ responses fall well above mere chance, indicating that the categorization of

stimuli in two distinct classes is still functioning, and hence does not necessarily require consciousness.¹⁰

Categorization is the main function of cortical visual neurons, in that each neuron is *feature-selective*: it only responds to a stimulus when that stimulus possesses certain visual features. A Nobel prize was awarded for this finding, as it is fundamental to the operation of the visual cortex (Hubel 1982). It ranges from low level features such as spatial frequency, orientation, direction of motion, or colour to higher-level features such as the geometry of a shape or the class of an object. Each feature-selective neuron can be seen as doing a simple, often one-dimensional categorization: it signals “vertical orientation”, “moving upwards”, or “rectangular shape” (Lamme & Roelfsema 2000). Face-selective neurons shout “face!” (Oram & Perrett 1992). The categorization responses of visual neurons are so fundamental to their operation that they are fully independent of consciousness: most neurons are equally feature selective in anaesthesia as they are in the awake condition (Dow et al. 1981; Snodderly & Gur 1995; Lamme et al. 1998a). Feature-selective responses of neurons are mediated via feedforward connections, and visual categorization proceeds along these feedforward connections in an unconscious way (Lamme et al. 1998b; Lamme & Roelfsema 2000).

Additional evidence dissociating categorization from consciousness comes from a multitude of sources. Unseen stimuli in backward-masking are also categorized, as can be judged from the specific priming effects they may evoke. For example, a fully masked digit 7 may speed up (or slow down) responses to categorizing a second digit (or number word) as either being above or below 5, showing that the masked and unseen number (the 7) is categorized according to its numeric value (Dehaene et al. 1998).¹¹ Many similar examples exist. Moreover, it has been shown that masked and hence unseen stimuli evoke category-specific responses from the brain, either in the form of se-

¹⁰ Note, however, that categorization is typically far better for stimuli than patients—or normal subjects—are aware of.

¹¹ Or more precisely: as being either below or above 5, in this experiment.

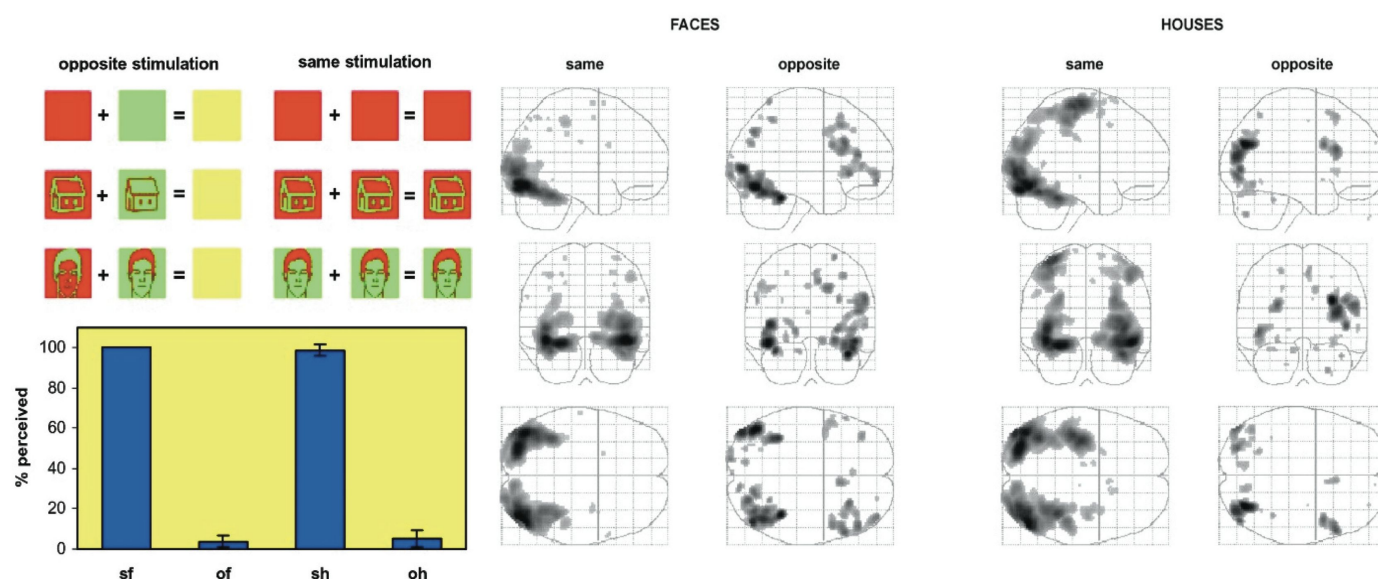


Figure 1: Faces and houses were made invisible using dichoptic masking—i.e., presenting oppositely coloured versions to each eye. Regardless of (in-)visibility, these faces and houses evoked selective activations of category specific regions of the brain (from: Moutoussis & Zeki 2002).

lective single unit responses (Rolls & Tovee 1994; Macknik & Livingstone 1998), or in the form of selective activation or category-selective regions such as the Fusiform Face Area (FFA) (figure 1) (Moutoussis & Zeki 2002; Kouider et al. 2009),¹² or in the Visual Word Form Area (Dehaene & Naccache 2001)—indicating that they are categorized up to the level of face vs non-face or word vs non-word (Dehaene et al. 2004). There is a large body of literature covering the unconscious processing of emotional valence in either faces or words (Straube et al. 2011).

Particularly far-reaching levels of unconscious categorization have been reported for behaviourally or socially relevant stimuli. Tools evoke selective activation of the dorsal stream

areas—and selective priming effects—when made invisible with CFS (Fang & He 2005; Almeida et al. 2008). Faces that have their eyes turned towards the viewer break from CFS sooner than faces that are turned away—a finding that is probably explained by the fact that faces turned towards the viewer pose a very relevant or even threatening social signal (Gobbini et al. 2013). Similarly, the gender of naked bodies is processed during CFS (Jiang et al. 2006). Also, the mismatch between object categories is identified for stimuli made invisible using CFS: scenes with mismatching objects (e.g., a cook taking a chess-board out of the oven instead of a dish) break from CFS sooner than matching scenes (Mudrik et al. 2011).

The latter finding is related to various non-visual “categorization” processes that occur for invisible stimuli: it has been shown that masked stimuli travel throughout the brain, even reaching high-level areas involved in inhibitory cognitive control, response error selection, or evidence accumulation, exerting high-level cognitive effects (Van Gaal & Lamme 2012). So invisible stimuli not only activate visual categorization processes, but also activate extremely high-level and very abstract categories such as the stimulus being a “stop signal”, an “error”, or “evidence for a right button press”.

¹² It is unclear to what level invisible faces are processed exactly. Clearly, face/non-face categorization takes place for masked stimuli (see below), but whether face identity is also preserved depends on the exact experiment. Some find face-identity-specific priming and suppression of activation of the FFA and related face-selective-regions for backward masked stimuli (Kouider et al. 2009). However, this effect was only present for famous faces, not for unknown faces, showing that it may not be identity itself that is processed but “level of fame” or something similar. Others have made faces invisible using CFS, and found that face-specific adaptation only occurred for visible, and not for invisible faces (Moradi et al. 2005). The two studies are hard to compare, partly because of the different techniques used to make faces invisible (masking vs. CFS), but mostly because the latter used an adaptation effect as independent variable. It may be that unconscious categorization still occurred, yet did not result in learning (e.g., Meuwese et al. 2013; Meuwese et al. 2014).

From a neural perspective, categorization is feature selectivity, which may range from very simple to highly complex features and categories. This kind of categorization proceeds entirely independent of consciousness.¹³ So how does conscious recognition differ from categorization? To answer this question, we have to take a closer look at categorization responses. What a face-selective cells does, is to categorize a face as belonging to the class of faces versus non-faces. That's all. When we consciously see a face, however, we do much more than this: we classify the stimulus as a face, but at the same we identify its shape, colour, identity, and emotional expression. So we distinguish between “that brown face of my sad-looking friend Peter” and very many other faces—and also between that face and millions of other potential visual stimuli.

Gulio Tononi uses the metaphor of a photo-diode to illustrate the point (2008, 2012). For a photo-diode a black screen is different from a white screen. That's a distinction it can make. The photo-diode carries information about the brightness of the screen, so its signal carries one bit of information (or a few bits, depending on its sensitivity). For us, however, consciously seeing a black screen is very different. Seeing the black screen implies that we distinguish it from a grey screen, a red screen, a black table, a green house, a pink face, a dog, a sound, a feeling, or any other sensory event that would have been possible. Consciously seeing the black screen thus carries a huge amount of information, as it excludes an almost endless set of alternative sensations. And that makes seeing “that brown face of my sad-looking friend Peter” very different from what a face-selective neuron does when it signals “face”. The neuron behaves much like the photodiode, in that it signals presence or absence of a feature along a single dimension. That is because neurons tend to combine feature-selectivity with invariance

for other features: a face-selective cell signals faces regardless of colour, size, identity or expression (Rolls 1992).¹⁴

Tononi proceeds from a photo-diode to the photo camera as a metaphor for explaining another central feature of conscious sensations (2004, 2008, 2012). He argues that the critical difference between a conscious representation in the human mind and what happens in a camera is that in the camera information is distributed and not integrated. Each and every pixel signals a particular level of luminosity, but it does so entirely on its own. It does not “know” what other pixels are doing. To the camera it would not matter if all the pixels were cut apart and became separate cameras. Conscious sensations, on the other hand, are integrated.

It thus seems that to find for visual operations that are more closely linked to consciousness, we must look for something beyond basic categorization. We must look at processes where the individual pixels in our camera—the billions of neurons each signalling particular features—are interacting, and are integrating their information.

5 Interference: A loss of independence

The pixels in the “camera of the human mind” do not work independently. A strong case in point are illusory brightness or colour shifts. Patches of the exact same brightness may be perceived as entirely different, depending on their surroundings, and depending on the global configuration of brightness and contrast. A striking example is the cylinder with checkerboard illusion shown in the right half of figure 2. Similar illusions exist in the domain of colour (figure 2, left). Relatedly, everyone who has ever tried to paint a picture has experienced that it takes an astonishingly rich palette of reds, purples, browns, yellows, and even greens or blues to construe a veridical depiction of a simple red apple. The unitary experience of see-

¹³ Another illustration of the separation between feature selectivity and conscious experience is the observation that many neurons signal features of which we are not aware: V1 neurons signal the orientation of gratings that are of too finely spaced for us to perceive (He et al. 1996; Foster et al. 1985), respond to 3D disparity where we do not see depth (Cumming & Parker 1997), or signal invisible temporal frequencies (such as the flickering of light beyond the flicker-fusion frequency of about 15–25Hz, Maier et al. 1987).

¹⁴ Responses are modulated by such features, but typically this happens only after some delay (Sugase et al. 1999). The initial feedforward response is typically fully governed by a basic feature, like face vs non-face. Later on, responses are modulated by face identity or expression, and this is mediated by horizontal or recurrent interactions between neurons. We then enter the domain of feature integration, which is a hallmark of conscious recognition; see below.

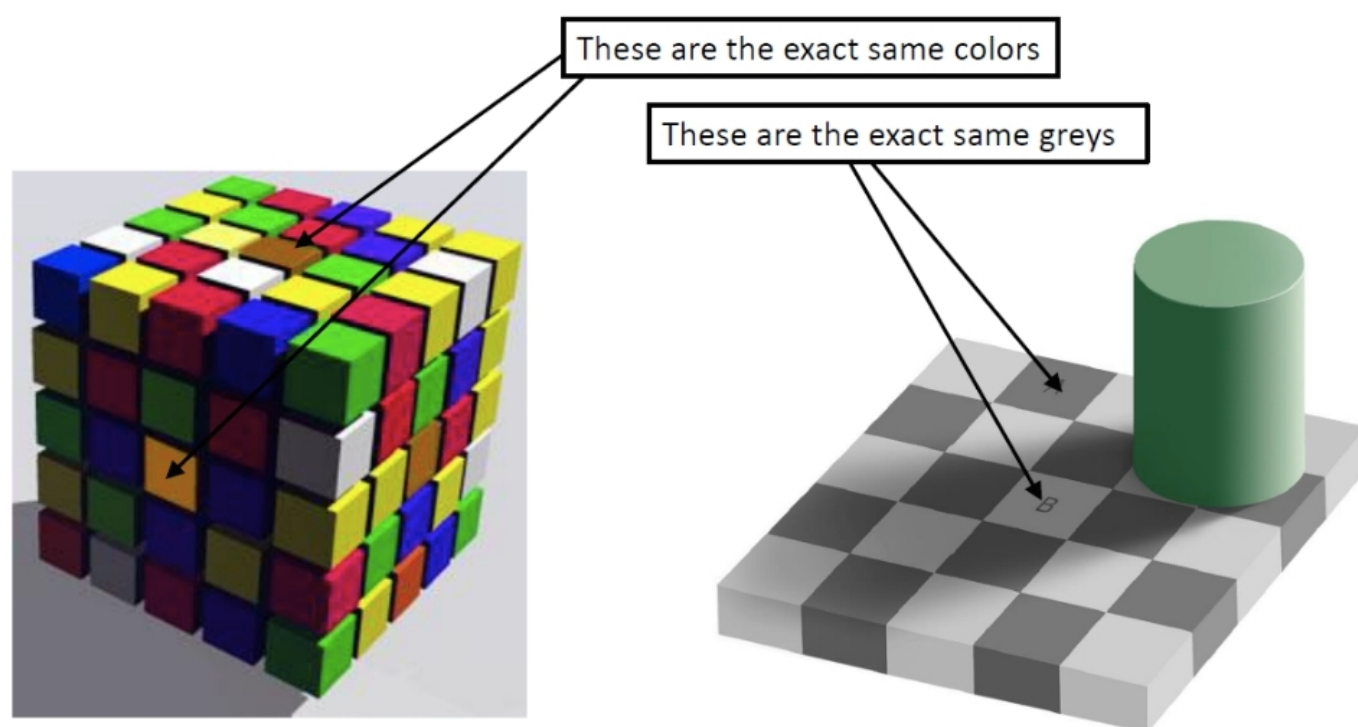


Figure 2: Two strong shifts in the perception of colour and brightness. Although the indicated patches are identical, they are perceived as having quite different colour and brightness. Visit Michael Bach’s website (<http://michael-bach.de/ot/>) for these and many other examples.

ing a red apple is in fact composed of the detection of a multitude of wavelengths, all interacting to compose that one colour. Only with extreme focused scrutiny (or by covering surrounding elements) are we able to isolate the elements that make up our unitary conscious experiences.

Another illustration is the phenomenon of colour constancy. When we look at a bowl of fruit in the blue morning light the spectral composition of wavelengths reflected from the fruits is very different from the wavelengths coming from the fruits at sunset (figure 3). Nevertheless, we see the banana or the apple as having the same colour whether it is dusk or dawn. Our visual system is not interested in the wavelength coming from fruits; it is interested in their potential taste or edibility. Therefore, it discounts the illumination, and computes “colour”, which is a property of the object, rather than of the light coming from it.¹⁵ Colour is not wavelength; colour is a meaningful property of

objects that is based on wavelengths, yet transcends it.

To what extent do these phenomena depend on consciousness? Harris et al. (2011) studied a brightness illusion much like that in figure 2. Two circles were shown, of either the same or different brightness. By placing these circles in a dark and bright surround respectively, their brightness suffered from an illusory shift. In the critical condition, the surrounds were made invisible by presenting them to one eye, and filling the other eye with a continuously flashing Mondrian stimulus. This resulted in CFS of the surrounds. Cleverly, the two circles were shown in both eyes, so remained visible throughout. Regardless of the CFS-induced invisibility of the surrounds, the circles still showed illusory brightness shifts.¹⁶

The neural mechanisms of illusory brightness perception were studied extensively

¹⁵ It probably discounts the illumination by very much the same mechanisms as the illusory brightness shifts discussed above (via inhibitory lateral interactions). However, precise neural mechanisms may be different, as might be the cortical level at which neural responses reflect colour rather than wavelength.

¹⁶ It must be noted that in this experiment, the surrounds were not always fully invisible. In 86% of the trials, subjects reported not seeing the surrounds. Only these trials were used for the analysis. Within these trials, discrimination of the background (is the darker half left or right?) was at chance level, leading to the argument that indeed there was a full absence of awareness of the surround.



Figure 3: These images show a bowl of fruit photographed in three lighting conditions—artificial light (left), hazy day-light (middle), and clear blue sky (right). Notice the marked variation in colour balance caused by the spectral properties of the illuminant. We are not normally aware of this variation because colour constancy mechanisms discount illumination effects (image and legend from <http://www.psypress.co.uk/mather/resources/topic.asp?topic=ch12-tp-04>).

in the macaque monkey and cat visual cortex. It was found that perceived brightness (modulated by flanking regions) influenced neural responses in area V1 of the cat, but not at earlier stages such as the LGN or the optic tract, thereby showing a gradual progression from physical brightness to perceptual brightness in the visual pathways (Rossi & Paradiso 1999; Rossi et al. 1996). Using the Cornsweet brightness illusion,¹⁷ it was found that in the monkey's visual cortex, V2 cells represents surface brightness whereas V1 cells do not, pushing the level at which perceived brightness is calculated somewhat higher (Roe et al. 2005). Either way, these results were recorded in anaesthetized animals, showing their independence from consciousness.

How the visual system goes from the detection of wavelength towards the representation of colour is still a topic of controversy. Initially, there was thought to be a modular progression from V1 cells encoding wavelength towards V4 cells encoding colour. That view was challenged by various findings showing that the responses of V1 cells are influenced by surrounding hues. The view that V4 is the “colour module” has also been challenged, in part by strong disagreement on the homology between monkey V4 and alleged human counterparts.¹⁸ Moreover, the coding of colour is intricately

linked to the coding of object shape, and hence can no longer be viewed as a simple “add-on” to our visual percept.¹⁹ It is now thought that the perception of colour depends on the interaction between neuronal groups, or is best understood as a population code (Shapley & Hawken 2011).

Given this controversy, it is difficult to know to what extent colour perception depends on consciousness. Many of the recordings in monkey visual cortex were performed in awake animals, some in anaesthetized animals (Shapley & Hawken 2011). A clear-cut difference in results between the two conditions is hard to establish. A remarkable finding is that blindsight patients report no conscious sensation of colour, yet may have spectral sensitivity curves that have a similar shape in the lesioned and intact hemi-fields (Stoerig & Cowey 1989). Spectral sensitivity is, however, mostly carried by wavelength. Similarly, patients with cortical colour blindness (achromatopsia) do not consciously perceive colour, yet can detect objects or patterns based on wavelength contrasts (Cowey & Heywood 1997). Colour constancy mechanisms, on the other hand, are absent in the lesioned hemi-field of blindsight patients (Barbur et al. 2004; Barbur & Spang 2008), and

¹⁷ In this illusion, two surfaces of identical brightness are perceived as having different brightness, because there is a contrast edge between them.

¹⁸ I am not even going to dare mentioning their names here.

¹⁹ The fact that black-and-white photography works so well, has led us to believe that colour is a feature that is “painted” onto objects, as a sort of extra, independent of any other feature. We are now coming around from this view. For example, to compute the colour of an object, the object's shape has to be taken into account, otherwise shadings would be misinterpreted. Object identity also influences colour perception: a brownish colour on a banana will be seen as more yellow than it would on a tomato.

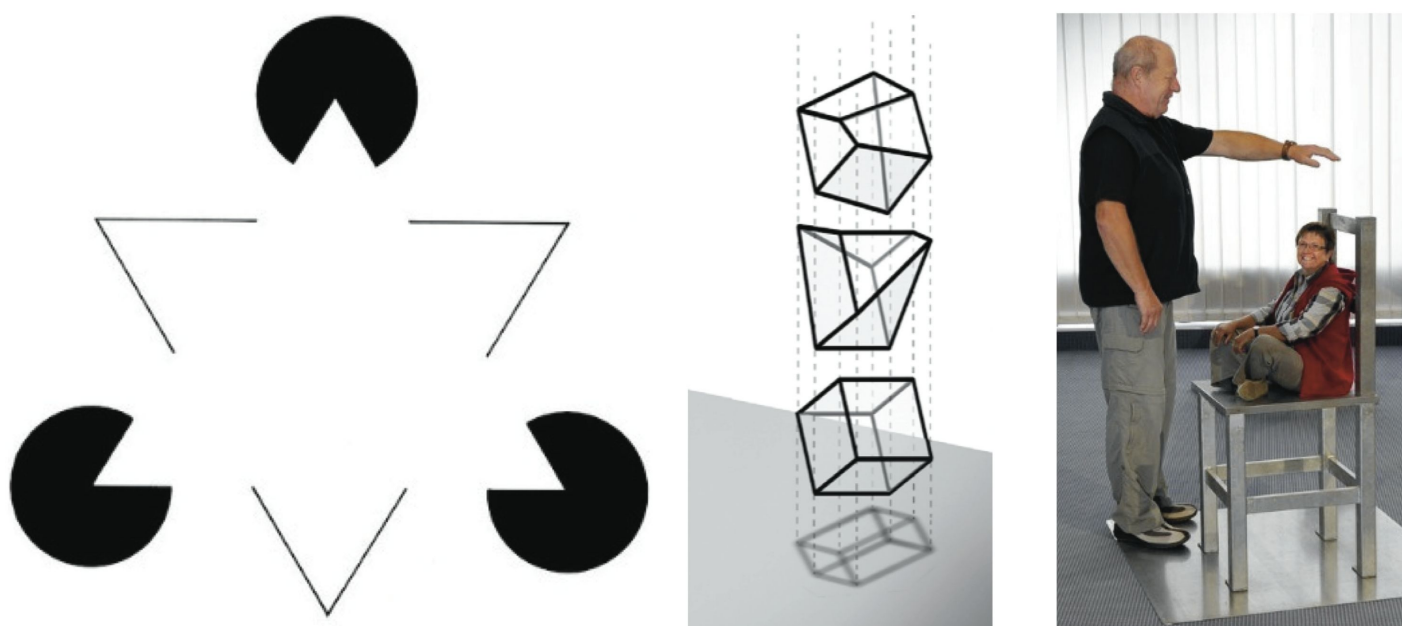


Figure 4: Left: the Kanizsa triangle. Note the illusory brightness increase inside the region of the illusory triangle. Middle: the 2D projection of a cube can in fact originate from a multitude of 3D objects. We regularly interpret it as a cube, however. Right: we see the woman as small, despite our cognitive ability to realize that “this cannot be true”. Our 3D “priors” force us to see her as small (from <https://richardwiseman.wordpress.com/2009/09/09/great-table-illusion/>).

hence seem more closely linked to conscious perception.²⁰

The difference between perceived colour and wavelength, and its relation to conscious vision, has been directly addressed in a masked priming experiment. In this experiment, subjects were shown desaturated blue, green, or white coloured disks. Perceptually, the white was closer to the blue disk. From the point of view of the phosphor activations on the monitor screen, on which the disks were shown (i.e., their “wavelength composition”), the white disk was, however, closer to the green disk. What was studied was the effects of these disks when they acted as primes for a subsequent colour discrimination. It was found that masked, and hence invisible white disks, acted more like green primes than like blue ones. Visible white disks, on the other hand, acted more like blue primes than like green ones (Breitmeyer et al. 2004; Breitmeyer et al. 2007). Apparently, unconscious priming acts on wavelength similarity,

whereas conscious priming acts on perceived colour similarity.

All in all, it remains difficult to assess the relation between consciousness and phenomena like brightness or perceived colour illusions, or mechanisms related to colour constancy. Perceived brightness seems to depend on largely unconscious mechanisms, and on fairly low level and short range mechanisms. The transition from wavelength analysis to the perception of colour is more likely to accompany the transition from unconscious processing to conscious vision. A firm conclusion, however, relies upon settling the debate about mechanisms of colour perception and their neural substrates in humans and animals, and more direct experimentation on how these mechanisms are affected by manipulations of consciousness.²¹

6 Inference: Beyond the input

In the phenomenon of colour constancy we have already seen a hint of another visual function.

²⁰ This argument is, however, weakened by the fact that other long-range colour interactions remain in the blind hemi-field (Barbur et al. 2004), and by the finding that colour constancy mechanisms may depend on fairly early, monocular mechanisms (Barbur & Spang 2008). Moreover, it is reckoned that several colour constancy mechanisms exist, some of which are based on retinal adaptation mechanisms (Kamermans et al. 1998).

²¹ Obviously, these empirical issues about colour perception and consciousness have very direct consequences for many philosophical debates as well, given the many thought experiments that rely on colour perception and the whole notion of qualia.

Colour is not about the wavelength coming from objects. It is a property of objects that we infer from wavelengths. At some point, conscious perception starts to diverge from the mere physical properties of the input, in a process we call *inference*. There are many more examples of inference, and many visual illusions capitalize on the fact that our visual mechanisms are constantly trying to make sense of the world. Figure 4 shows the famous Kanizsa triangle. The minimal, strictly physical interpretation of the image is that of three Pac men pointed at each other and three arrowheads pointing outwards. But our perceptual interpretation goes beyond this, in that we *see* a white triangle hovering over three black circles, occluding another outlined triangle. The illusory triangle is seen as slightly brighter than its surround, and illusory contours mark its “borders”.

This process of inference seems to strongly fit the intuitive difference between a camera and conscious vision. It requires the integration of multiple “pixels”, their interaction, and their interpretation beyond what is strictly given by the image itself. And it is in this last aspect in particular that prior knowledge about the world comes into play, and starts to interfere with the stimulus-driven feature-selective categorization of the input.

The Kanizsa triangle can be seen as a specific example of the more general propensity of the visual system to arrive at a representation of surfaces in 3D space (also called the 2.5D sketch). In that representation we seek the most natural interpretation, consistent with our existing experience of how things are in the world. It is simply much more likely that there is a triangle covering circles than that there are three Pac men that happen to be facing each other at exactly 60° angles. The triangle interpretation is generic, whereas the Pac men one would be accidental (Albert & Hoffman 2000). Nakayama & Shimojo (1992) studied various configurations of 3D stimuli, and found that our visual system always strives towards the interpretation that is most generic, i.e., that would least depend on an accidental viewing position of the observer. Interpretations that would not change when the observer happened to shift position are fa-

voured, given that we are constantly moving relative to objects. For example, the 2D image of a cube can in fact arise from an infinite number of shapes (figure 4, middle), yet we tend to favour the “cube” interpretation because it is the most generic one.

Another way of putting it would be to say that the cube interpretation fits our common experience, in that most of the time, these kinds of 2D projections arise from regular 3D cubes: it is the most ecologically valid interpretation. In a modern guise, this aspect of inference is formalized as a Bayesian approach, where vision uses a set of prior probabilities to arrive at the most likely 3D interpretation of a 2D image. The cube has the highest prior, compared to the more irregular shapes. Illusions like the Ames room (where someone changes size when he walks from one corner to the other), or the size illusion shown in figure 4 (right) capitalize on these assumptions: we assume that rooms have rectangular floors and walls, we assume the woman is sitting on a chair. These assumptions are so strongly embedded in our visual hardware that even in the face of strange consequences, such as people growing in size within a few steps or a man holding his hand over a mini-woman, this inference is maintained.

Many more illusions display non-veridical inferences. In the Ebbinghaus illusion, the perceived size of a disk depends on the size of surrounding disks. In the Ponzo and Müller-Lyer illusions we see line segments as having different lengths, while in fact they are the same. These illusions show that the size of an object is an inference that we draw from its context, rather than from the space it occupies on the retina.

To what extent does inference depend on conscious vision? When we have to pick up the disks in the Ebbinghaus illusion, it appears that our hands open at a pre-grip aperture that is in accordance with the disk’s actual size, not its perceptual size. Apparently, size context effects influence perception, and not automatic action—which has led to the idea that we have two largely separate neural pathways, one transforming visual input into conscious perception, and the other translating visual input into automatically guided action (Goodale & Milner 1992).

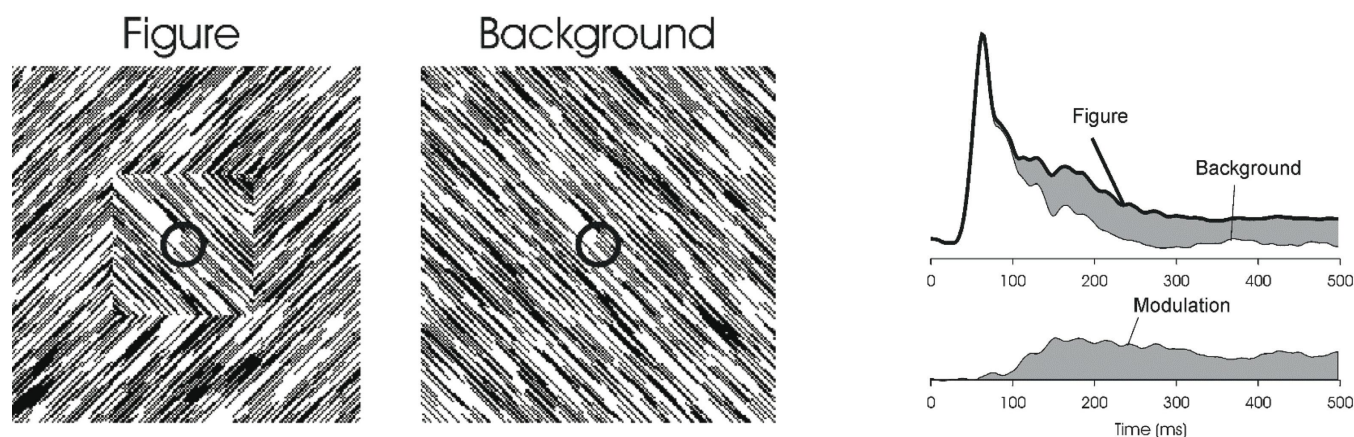


Figure 5: On the left, we see a textured square overlying a textured background. This is because we automatically group all line segments with the same orientation into one object, and segregate it from the line segments with the orthogonal orientation. The small circle represents the receptive field of a V1 neuron, that would not be able to differentiate between the “figure” and the “background” stimulus, because the line segments within that receptive field are identical. Indeed, V1 responses are identical up to about 100ms after stimulus onset. Beyond that, the two responses start to diverge, however, indicating that the response of the V1 neuron is modulated by the perceptual context of what is within its receptive field (Lamme 1995; Lamme et al. 2000).

There is more evidence linking perceptual inference to conscious vision. Harris (Harris et al. 2011) studied whether Kanizsa triangles were still inferred when the inducers were made invisible using CFS. The same setup was used that showed the presence of brightness illusions under CFS (see above). Subjects had to indicate whether the triangle in the suppressed eye was pointing left or right. They were at chance level, indicating that the Kanizsa-type inference depends on consciousness. Another study, however, found that Kanizsa triangles broke through CFS earlier than control stimuli with inducers pointed outwards (Wang et al. 2012), suggesting that Kanizsa-type inference does occur pre-consciously.

At the single neuron level, the detection of illusory contours has been studied quite extensively. Initially, it was found that V2 cells respond in an orientation-selective manner to Kanizsa-type illusory contours (Von der Heydt et al. 1984). More recently, other areas have been shown to be involved as well (Sáry et al. 2008)—area V4 in particular (Cox et al. 2013). And in human neuroimaging studies it was found that Kanizsa-type illusory contours activate many early visual areas (Seghier & Vuilleumier 2006). All these studies used awake animals or humans, so it is difficult to

infer whether these responses depend on the conscious state.

Marcel studied the processing of illusory triangles in two blindsight patients. Two inducers were presented in the sighted hemi-field, while one critical inducer was presented in the blind field, either completing the triangle or not. Completed triangles were detected far above chance (~80%), while the detection of the inducer shape was at chance. Moreover, one of the subjects described the illusory triangles as “brighter”, “out there on the screen” and “on top of something” (Marcel 1998).

All in all, the relation between inference and consciousness is unclear, mostly because fairly little work has been done as yet to study the relation directly (i.e., to study the effect of consciousness manipulations on inference and its neural correlate), but also because much of the work that has been done focuses on a single (though very important) phenomenon: the Kanizsa triangle.

7 Integration: Feature grouping and segregation

Both in interference phenomena such as brightness or colour shifts and in inference phenomena like the Kanizsa triangle we see some

aspects of the integration of information. Visual responses go beyond the encoding of individual pixels, and start to influence each other, either on the basis of more or less hardwired lateral interactions, or on the basis of the incorporation of prior knowledge. In the end, conscious vision seems to be about reaching *full* integration.²² We have one visual percept, where all information is combined.²³ This is a property of conscious vision that has interested scientists for a long time. Gestalt psychologists formulated a multitude of laws, along which image elements may be combined into larger wholes (Rock & Palmer 1990; Wagemans et al. 2012). In this grouping process, all features, together with their interactions, inferences, and meanings are combined into a final percept: the thing we see, the whole scene containing shapes, objects, and backgrounds. This is a highly dynamic process in which various Gestalt laws may compete for one interpretation or another, and where subtle changes may influence the meaning of pixels at long distances. We enter the domain of feature integration, grouping, binding and segregation. In short; the domain of perceptual organization.

Two levels of integration may be distinguished, where a subdivision between “base groupings” and “incremental groupings” may be useful (Roelfsema 2006). Base groupings are those that depend on the fact that some feature combinations automatically ride together. An orientation-selective cell in the primary visual cortex, for example, is often at the same time

also direction-selective. It may be tuned to particular binocular disparities as well. And it will have a limited receptive field. So the firing of that neuron already goes beyond a one-dimensional feature-detector, beyond the photo-diode. It signals an orientation, moving in a particular direction, at a particular 3D depth, and located in some part of the visual field. Such base groupings exist for many feature combinations, such as colour and shape (e.g., V4 cells), or motion and disparity (e.g., middle temporal, MT, cells).

Another type of base grouping is visible in the feature selectivity of a particular cell, where we may recognize the combination of feature-selectivity of cells at earlier levels. From the start, Hubel and Wiesel recognized that orientation selectivity could be viewed as a convergence of information from retinal ganglion cells lying in a row. The feedforward convergence of information from orientation selective simple cells leads to the receptive field structure of complex cells, which are orientation and direction selective (Hubel & Wiesel 1968). Many higher-level feature-selective cells can be seen as converging information from lower level cells (Tanaka 1996).

Base grouping does not depend on consciousness. The combined feature selectivity of neurons, as well as high-level feature selectivity based on the feedforward convergence of lower-level feature selectivity are still present in anaesthesia or masking (Lamme & Roelfsema 2000; Roelfsema 2006).

Of a very different nature are “incremental groupings”. Here, the information from separate neurons has to be combined to obtain a higher level categorization. A good example is texture based figure-ground segregation, shown in figure 5 (Lamme 1995; Zipser et al. 1996). Here, we automatically perceive a textured square overlying a textured background. This is entirely due to the fact that the centre square is made up of line segments of a particular orientation, different from the line segments that make up the background. There is no luminance difference or any other cue that gives the square “away”. Line segments of one orientation are automatically grouped into a coherent surface—the square—that is segregated from the surface that is

²² Tononi similarly argues that consciousness always strives for “maxima of integrated information”, for which he uses the metaphor of the internet (2012). Like the brain, the internet is a highly interconnected structure where information travels from one part to other parts. In contrast to the brain, however, the internet is designed to transfer information from one specific part of the net (computer A) to a specified other part of the net (computer B), and it would in fact be rather counterproductive if this information were influenced by other information flowing from computers C to D or E to F. At another moment information may flow from A to C or D or F. The internet therefore does not strive for “maxima of integrated information”, whereas the brain typically does. Focussed attention, in such a view, would then be in fact a mechanism that counters this propensity towards maximally-integrated information, and which enables the brain to operate more strongly along the principles of the internet.

²³ This is in fact such a strong intuition that it has led us to believe for a long time that consciousness must be some place in the brain “where it all comes together”. Descartes envisaged the pineal gland as such a place, and hence theories that lean towards such an explanation of consciousness are often said to suffer from the fallacy of the “Cartesian theatre”.

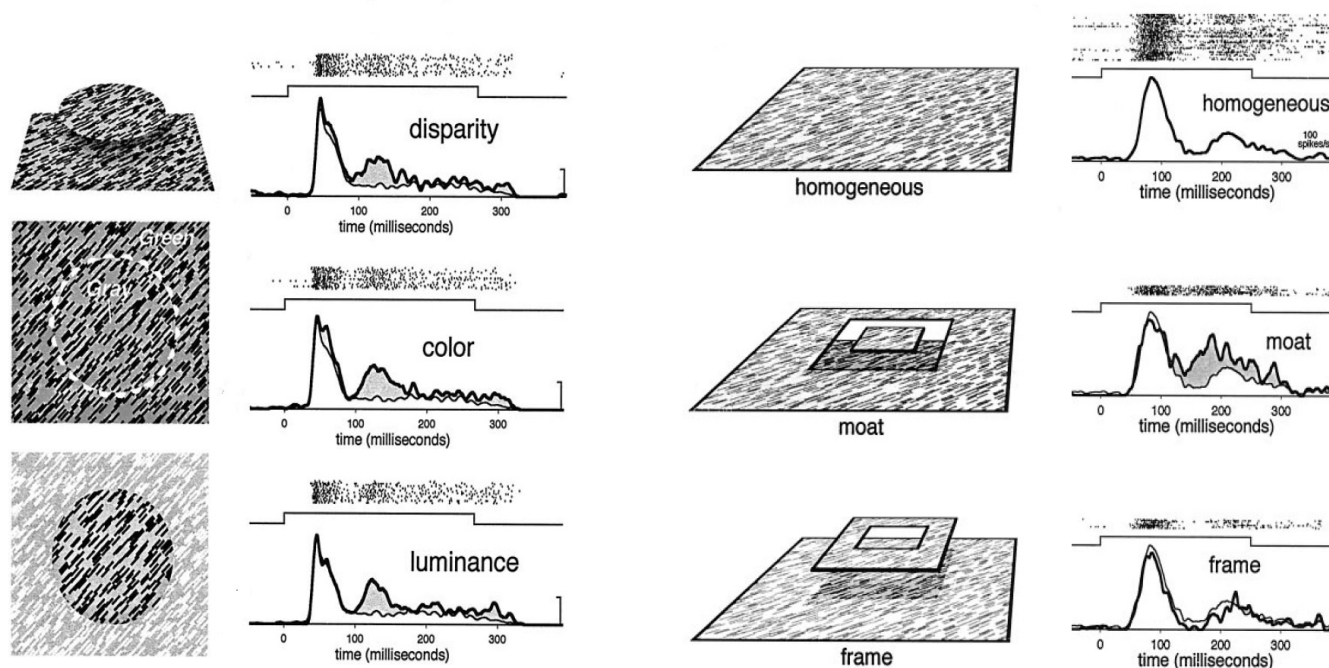


Figure 6: Contextual modulation of V1 responses follows the global perceptual interpretation of images. In all cases, the V1 receptive field is stimulated with the exact same line segments. When these line segments belong to a homogeneous background texture, a response indicated by the thin line is given. Left: when the line segments belong to a figure that is defined by differences in disparity, colour, or luminance, the responses are larger. Right: differences in 3D disparity were used so that the patch of texture was either part of a figure square “floating in a moat behind it” or in the background with a “frame” hovering in front of it. The contextual modulation always followed these perceptual interpretations, in that “figure” interpretations always evoked larger responses (Zipser et al. 1996).

formed by line segments of the other orientation—the background. Orientation-selective neurons in V1 typically have small receptive fields, which would only cover a small part of either the figure or background. The grouping of line segments into coherent surfaces, segregating from each other, requires the integration of information from a large set of separate V1 cells. This constitutes “incremental grouping” (Lamme & Roelfsema 2000; Roelfsema 2006).

The neural basis of the integration of image elements into larger units, and the subsequent segregation of such units into figure and ground has been studied extensively at the single unit level, both in anesthetized and awake monkeys. The key finding is that of “contextual modulation”, where the response of a neuron to a particular feature within its receptive field is modulated by the larger perceptual context of that feature (Lamme 1995; Zipser et al. 1996; Lamme et al. 1999). In the example of figure 5, the small circle represents a V1 receptive field,

which typically has a size of ~ 1 degree of visual angle. From the “point of view” of that receptive field, there is no difference between the “figure” or the “background” stimulus: in both cases, identical line segments cover the receptive field, and if the neuron were just signalling the presence of this feature (“left diagonal orientation present”), the responses of this neuron should be identical for the two stimuli. Indeed they are, as shown in the panel on the right, showing fully overlapping responses, until ~ 100 ms after stimulus onset. At that point, however, the responses for figure and background start to diverge. Apparently, information on the context of the line segments starts to influence the response, so that the response is larger for the “figure” than for the “background” context (Lamme 1995).

These kinds of figure-ground modulations follow the perceptual interpretation of scenes to a large extent. For example, when figure-ground relationships are ambiguous, or reversed, the



Figure 7: Contour grouping. In all cases shown here, oriented image elements are grouped together to form either a line (left), a circle (center), or an animal (right). They group according to the Gestalt principles of proximity, similarity, and colinearity. These stimuli were also used in neurophysiological experiments, typically showing that elements that group and segregate evoke larger neural responses than isolated or background elements.

modulation follows the globally-organized percept, rather than local orientation differences or gradients (figure 6, right panel) (Zipser et al. 1996; Lamme & Spekreijse 2000).

The perceptual grouping of image elements into larger units follows certain rules and principles, the formulation of which was the largest contribution of the Gestalt psychologists to modern vision theory (Wagemans et al. 2012). Among these Gestalt laws of perceptual organization are “similarity” (elements that look alike will be grouped), “common fate” (elements that go together in time, e.g., move together, will be grouped), “proximity” (elements that are close together will be grouped), and “good continuation” (elements that lie along a smooth line will be grouped). Contextual modulation of V1 neurons behaves according to these rules, in that elements that share luminance, colour, disparity, orientation, direction of motion, or co-linearity induce facilitatory interactions (figure 6 & 7) (Lamme et al. 1993; Lamme 1995; Kapadia et al. 1995; Zipser et al. 1996; Lamme et al. 2000).

How does Gestalt grouping and segregation depend on consciousness? To some extent, contextual modulation seems to survive during anaesthesia. This is, however, largely limited to fairly short range interactions between neurons, barely beyond or entirely within the receptive field (Allman et al. 1985; Gilbert & Wiesel

1992; Nothdurft et al. 1999). More long-range interactions, and interactions that express more global scene interpretations can only be recorded in awake monkeys (Knierim & Van Essen 1992; Lamme 1995; Kapadia et al. 1995; Zipser et al. 1996). For example, the figure-ground specific modulation of V1 responses shown in figures 5 and 6 (and structure from motion defined figure-ground modulation) is fully absent when monkeys are anaesthetized. At the same time, the orientation and motion selectivity of these neurons (i.e., their ability to categorize certain features) is not affected at all (Lamme et al. 1998a).

Similarly, backward masking disrupts figure-ground modulation. In monkeys, the visibility of texture orientation defined figure-ground targets was manipulated by masking with a stimulus consisting of randomly-positioned texture-defined figures (figure 8). The animals were at chance in detecting the location of the target figure for stimulus-onset asynchronie (SOA) of up to 50 ms (i.e., 50 ms between the onset of the target figure and the mask). At larger SOA's, behaviour quickly rose to ceiling. Figure-ground contextual modulation followed the same pattern: absent up to and including SOA's of 50 ms, and increasingly present at longer latencies. At the shorter latencies, however, V1 neurons still responded vigorously to the texture patterns in an orientation-selective manner,

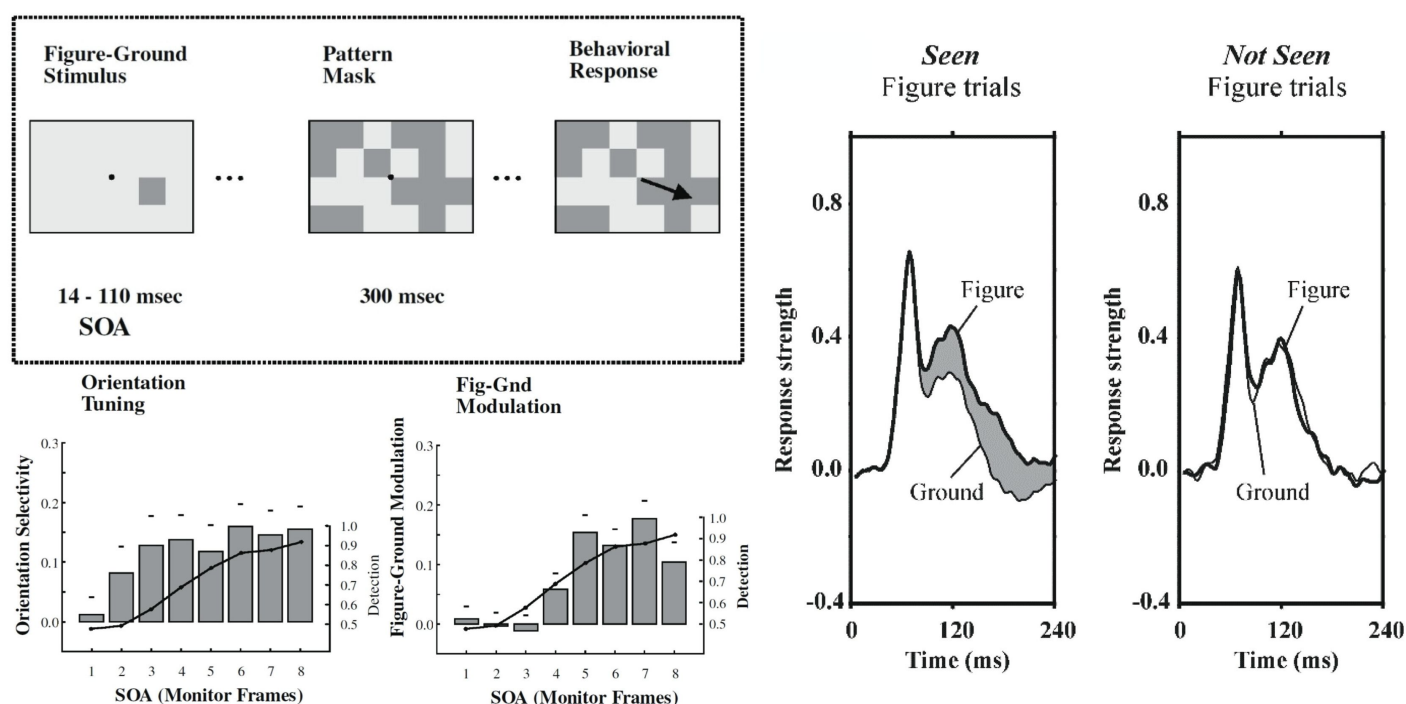


Figure 8: Left, above: textured figure-ground squares (like shown in figure 5) were presented either left or right of the fixation spot, and monkeys had to indicate their position with an eye movement. The figure targets were masked with a pattern of randomly-positioned texture squares. Left, below: the graphs show—for different SOA's—the ability of monkeys to correctly identify the position of the squares (line graph) versus the strength of either orientation-selective responses or figure-ground modulation (bars). Monkeys do not see the figures at SOA's of up to 3 frames (~50ms), and likewise, contextual modulation is absent in those cases, whereas orientation selectivity is not (Lamme et al. 2002). Right: monkeys had to indicate the presence or absence of textured figure targets by making an eye movement or deliberately maintaining fixation. When figures were not seen, contextual modulation was absent (Supér et al. 2001).

showing that lower level classification was still present for unseen orientations (Lamme et al. 2002). Similar results were obtained in human subjects using EEG responses (Fahrenfort et al. 2007).

Contour grouping, as displayed in figure 7, is particularly susceptible to masking. When these displays are temporally alternated, so that each element rotates 90° in successive displays, a strong masking effect is observed.²⁴ Depending on the angle between elements forming a contour, visibility drops to chance at alternation frequencies between 12 and 1Hz. This implies that the integration of these contours takes between 80 to 1000 ms (Hess et al. 2001).

Zipser used dichoptic masking to render orientation-defined figures invisible. Figure-

ground stimuli like those of figure 5 were shown to the two eyes of awake and fixating monkeys, yet with opposite orientations in either eye. As a result, the dichoptically-fused image consisted of cross-like elements, in which a figure was no longer visible.²⁵ Figure-ground modulation was absent in this case (Zipser et al. 1996). In a similar experiment in human subjects, Fahrenfort used face stimuli that were defined by oriented texture differences. A face was present in each image presented to the two eyes. Yet when binocularly combined, the face disappeared in the fused percept. He compared the neural signals obtained for such stimuli to responses to similar stimuli where binocular fusion resulted in a vis-

²⁴ This manipulation is a combination of backward and forward masking, and also somewhat reminiscent of dichoptic masking, in that in subsequent displays images with the opposite orientation contrast are shown. See the two images of figure 9, but then not presented to the two eyes but in rapid alternation.

²⁵ A similar setup was used in the curious case of alleged "blindsight in normal observers". In one of the experiments in that paper, target figures were made invisible using the same manipulation of dichoptic presentation of orthogonally-oriented elements. It was claimed that despite their subjective invisibility, subjects were able to localize the targets above chance, just as blindsight patients do for unseen stimuli (Kolb & Braun 1995). The findings were not replicated, however (Robichaud & Stelmach 2003).

ible face (figure 9) (Fahrenfort et al. 2012). A striking finding was that visibility (although rigorously checked behaviourally) had no effect on the ability of the Fusiform face area to distinguish between face and non-face stimuli, once more corroborating the independence of categorization responses and consciousness. In addition, invisible face stimuli could be classified from neural responses when training the classifier on visible stimuli and vice versa. The difference between visible and invisible binocular faces was found in the fact that visible faces evoked strong recurrent interactions between the FFA and earlier visual areas, both expressed in the fMRI signal (assessed using psychophysiological interaction analysis with the FFA as a seed), as well as in the EEG signal (showing a larger amount of theta, beta and gamma synchronization, and the presence of figure-ground modulation only in the visible condition).

The most direct relation between contextual modulation and consciousness was perhaps demonstrated by Supér et al. (2001). Monkeys were shown oriented texture figure-ground targets at different locations, and had to signal their presence by making an eye movement towards their positions. Importantly, however, in 20% of the trials, no figure was presented at all, and the monkeys had to maintain fixation on those catch trials for the duration of the stimulus.²⁶ Indeed the monkeys refrained from making eye movements on catch trials (as they were trained to do). But also on some 8% of trials in which a figure *was* presented they maintained fixation, as if to say “I did not see a stimulus here”. There was a striking difference in the level of contextual modulation for seen versus not-seen figure targets: modulation was fully absent for not-seen figures (figure 8). Seemingly, on some trials contextual interactions spontan-

eously fail to develop, and the result is that figure targets were invisible.²⁷

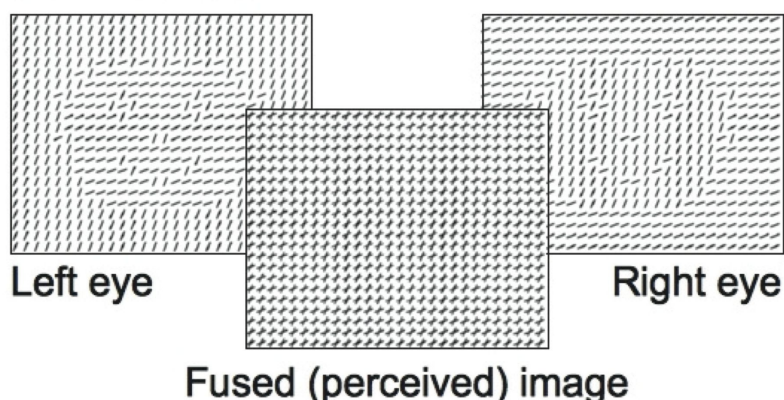
That brings us to the question of neural mechanisms. Seemingly, the visual functions of perceptual organization, grouping according to Gestalt laws, and figure-ground segregation all depend strongly on the conscious state, and on the objective (or subjective) visibility and perceptual interpretation of the stimulus. Do these functions have similar neural mechanisms? There has been much debate on the neural connections underlying contextual modulation effects. Given the latency of the effects (typically several milliseconds after the initial categorization or feature response) it was originally hypothesized that they depended on feedback signals from higher-level visual areas (e.g., V4, IT, MT, etc.) toward lower levels (e.g., V1, Zipser et al. 1996). Experiments using cooling or lesioning of higher-level areas gave mixed results. Local inactivation of V2 using GABA injections had no effect on short- to medium-range contextual effects in V1 (Hupé et al. 2001). Cooling area V5/MT, on the other hand, had effects on figure-ground signals in V1, V2, and V3 (Hupé et al. 1998). These effects, however, worked on the early part of the response, and were evoked using stimuli where segregation depended more on contrast differences than on the long-range integration of information (Bullier et al. 2001). Others also found figure-ground effects that were faster than those discussed here (Sugihara et al. 2011). There is thus a whole range of contextual effects, some of which are faster than others, and some of which may depend on feedback while others do not.

There is one counterintuitive aspect of interpreting these results in this way: in fact, feedback connections are not slow, but just as fast as feedforward connections, where both are at about 3.5 m/s (Girard et al. 2001). Horizontal connections that run via unmyelinated fibres in layers 2 and 3 of the cortex are about 10 times slower (Sugihara et al. 2011). Many of the Gestalt principles of perceptual organization

²⁶ This paradigm has been shown to distinguish between seen and not-seen stimuli in monkeys with a V1 lesion in one hemi-field, and was used to differentiate between “conscious” visual responses and unconscious blindsight behaviour: without catch trials (i.e., when in forced choice mode), monkeys react to both stimuli in the intact and in the lesioned field, expressing blindsight capabilities. In catch trials, however, monkeys only respond to stimuli in the intact and not in the lesioned hemi-field, as if expressing conscious sensation instead of a mere reflex (Moore et al. 1995). Supér et al. used the same paradigm in intact monkeys to assess conscious percepts of figure-ground stimuli.

²⁷ A later investigation into neural activity preceding either seen or not-seen figure trials showed that not-seen trials are preceded by somewhat lower level of spontaneous activity, and also express less inter-neuronal synchrony (Supér et al. 2003; Van der Togt et al. 2006).

Invisible fusion



Faces > Houses & Nonsense

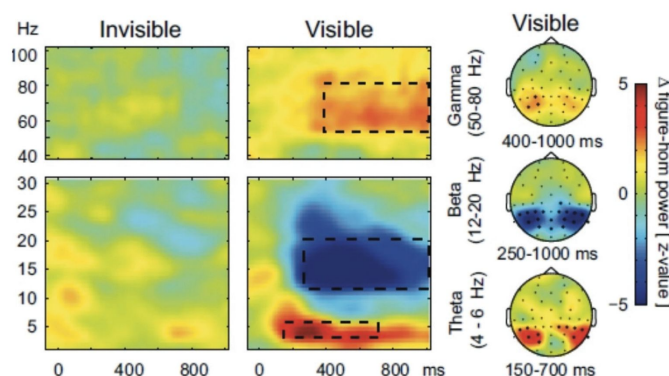
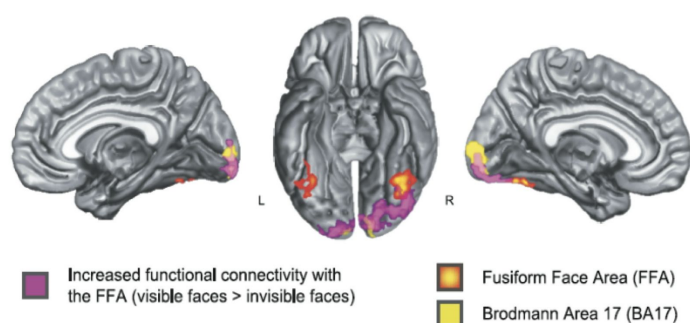
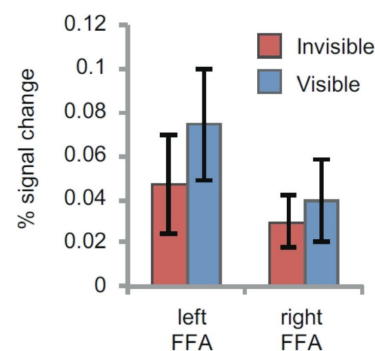


Figure 9: Top left: texture-defined faces were presented in either eye of subjects, yet with different orientations of line segments. As a result, the face was not visible in the fused percept (compare manipulation of figure 1). By using other orientation combinations, the same design could also result in a visible face (not shown). Top right: category-specific responses in the FFA did not differ for visible or invisible faces. Below: visible faces are characterized by strong recurrent interactions between FFA and earlier visual areas (left), and by strong synchronous activity in the theta, beta, and gamma bands (right). From: [Fahrenfort et al. \(2012\)](#).

are, however, embedded in these slow horizontal connections: V1 cells with a similar orientation preference are selectively interconnected via so-called patchy horizontal fibres. Moreover, these interconnections are strongest for oriented cells that have their receptive fields aligned along their orientation preference. Horizontal connections are also strongest between nearby cells ([Gilbert & Wiesel 1989](#); [Malach et al. 1993](#); [Bosking et al. 1997](#)). As such, these horizontal connections thus form the neural substrate of the well known Gestalt rules of “similarity”, “collinearity”, and “proximity”. A similar arrangement of preferred interconnectivity has been found for motion-direction selective cells in MT ([Ahmed et al. 2012](#)), potentially forming the substrate of the grouping principle of “common fate”. Neurophysiological correlates of these

grouping principles are relatively fast, however ([Knierim & Van Essen 1992](#); [Kapadia et al. 1995](#)).

The figure-ground segregation effects of figures 5, 6, 7, and 8 are among the longest latency contextual effects reported. That may be because they depend on both horizontal and feedback connections. Figure 10 shows the result of an experiment where the complete peristriate belt of visual cortex surrounding V1 and V2 was subjected to suction lesioning, removing (parts of) areas V3, V3A, V4, V4t, MT, MST, FST, PM, DP, and 7a ([Lamme et al. 1998b](#); [Supér & Lamme 2007](#)). Before the lesion, an oriented texture figure-ground stimulus evoked elevated activity in all neurons responding to the figure elements. Response modulation was even somewhat stronger, and occurred earlier at

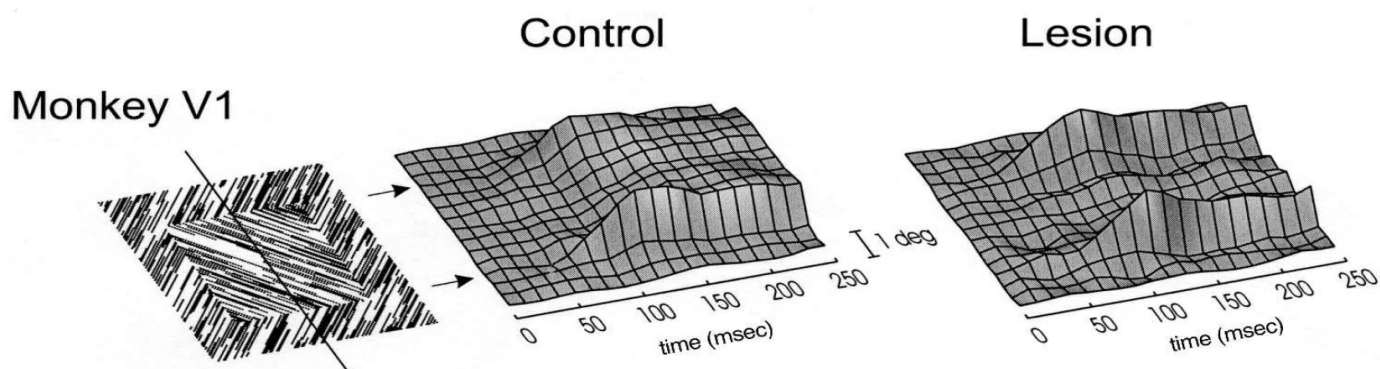







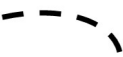

Figure 10: Contextual modulation (i.e., figure-ground responses, see figure 5) for various positions of the receptive field of V1 neurons (vertical axis), and extending over time (horizontal axis). In an intact monkey, modulation arises first at the figure-ground boundary, followed by a “filling-in” of the boundaries. After a lesion to the peri-striate belt of the visual cortex, only the boundary modulation remains, while filling-in has been abolished (Lamme et al. 1998a).

the boundary between figure and background. This was followed by a sort of “filling in” of enhanced activity between the boundary regions. We thus see an incremental process, starting with boundary segmentation and followed by surface segmentation. Similar findings have been reported in humans using combined EEG and TMS (Wokke et al. 2012).

After the lesion, the boundary enhancement remained, which may indicate that texture boundary detection mechanisms do not depend on feedback from higher visual areas and hence are mediated by horizontal connections within V1, or by recurrent interaction with V2. The centre modulation, where the figure elements are “neurally elevated” from the background elements, was completely abolished after the lesion, indicating that these figure-ground signals do depend on recurrent interactions between V1 and higher-tier areas. This finding was modelled on a realistic neural network of spiking neurons, indeed formalizing the idea that local orientation contrast—and hence the boundary between figure and ground—is mediated by inhibitory horizontal interactions between oriented receptive fields, whereas the figure-ground signal depended on excitatory feedback interactions trickling down from higher to lower areas (Roelfsema et al. 2002). Recently, laminar recording of figure-ground signals in V1 confirmed this idea (Self et al. 2013). These results show

that the long-latency figure-ground segregation effects depend on incremental interactions mediated by both horizontal and feedback connections. That may be the reason why they are most vulnerable to anaesthesia, masking, and other manipulations of consciousness.

Tononi modelled several neural architectures in order to find the connection parameters that fulfil the requirements for achieving maximally-integrated information. The optimal architecture consists of neurons that each have specific and different connections patterns, yet are sufficiently interconnected for each neuron to be able to connect to another via a few steps. Uniformly, or strictly modularly organized networks are less optimal. The thalamo-cortical system fits these requirements very well. On the one hand, neurons should be interconnected, otherwise information is not integrated. On the other hand, too much interconnection leads to a loss of specific information, as all neurons start doing the same thing, which happens in epilepsy or deep sleep—states that are indeed accompanied by a loss of consciousness (Tononi 2004, 2008, 2012). The contextual modulations that have been explored here seem to exactly express these properties: on the one hand, the neural responses are very specific, in that the major part of the response is driven by the features that are within the (small) receptive field. But on the other hand, the integration of these

VISUAL FUNCTION	Example	Conscious Vision	Anaesthesia	Hemianopia Blindsight	Backward Masking	Dichoptic Masking	Continuous Flash Suppression
Categorization Feature detection							
Higher level Categorization							
Interference					Brightness Colour		Brightness Colour
Inference				With intact hemifield			Breakthrough Discrimination
Base Grouping							
Incremental Gestalt Grouping			Short range Long range				
Figure-Ground Organization							

Function is Present
 Conflicting Results
 Function is Absent
 Unknown

Figure 11: Table summarizing the influence of consciousness manipulation on various visual functions. Colours indicate whether functions (rows) still operate under a particular manipulation (columns). In the case of conflicting or uncertain evidence (yellow), the cases or conditions where the function still seems present is written in green; the cases where the function is absent are written in red. All functions are assumed to be present in conscious vision. For each visual function, an icon depicts its most prominent example. See text for explanation.

features ride on top of that response as a moderate modulation, expressing perceptual integration that may cover a large spatial extent, yet never even beginning to fully override the information carried by the neuron. In other words, visual neurons have categorization as their main priority, yet they also integrate these categories at some point in their response. That is the moment in time where the seed for conscious perception is laid (Lamme 2003, 2006, 2010a, 2010b).

8 Is there a functional boundary between unconscious and conscious vision?

I have taken the two extreme ends of consciousness manipulation: clear-cut visible and above-threshold items in awake subjects veridically reporting their visual experiences versus visual processing in anaesthesia, blindsight, or during profound masking or suppression (figure 11). If we don't accept conscious vision in the former, and the absence of it in the latter, there is no use arguing about the phenomenon of con-

sciousness. Even so, it has been surprisingly hard to find fundamental differences in the workings of many visual functions in the two conditions. Categorization of visual stimuli, even up to high levels, clearly stands independent of conscious visibility. It is unclear whether interference—i.e., the fact that features are no longer treated independently—depends on consciousness: shifts in brightness perception do not depend on consciousness, while it is uncertain whether the transition from wavelength to colour perception (and colour constancy) marks the conscious-unconscious divide.²⁸ Similarly, the status of inference phenomena like those observed in the Kanizsa triangle is uncertain.

This is all the more surprising given that many of these functions have traditionally been viewed as expressing the transition from merely physical features detected by sensor arrays towards the perceptual interpretation of this in-

²⁸ Yet this transition may be the “Holy Grail” for those willing to understand qualia—or at least for those believing in “soft qualia”, i.e., phenomenal properties that are not entirely detached from visual functioning, and having some sort of neural substrate (Block 1996, 2005, 2007).

formation. Moreover, they mark the integration of stimulus-driven input with our knowledge of the world, such that we arrive at visual “meaning”. Recently, there has been quite some interest in so called predictive coding frameworks of vision (Rao & Ballard 1999; Panichello et al. 2012). In these frameworks, vision is seen as a type of Bayesian inference, where our prediction (prior) of the outside world is continuously matched with our sensory input, and where the difference is propagated through the network as an “error signal”, which then results in an updating of the model (posterior). Indeed, expectations bias our perception of the world, most strongly in the face of ambiguous stimuli, but also in the case of unambiguous stimuli (Panichello et al. 2012). Although it has been suggested that either the matching process, the prior, or the posterior in this type of inference have some relation to consciousness,²⁹ this is questionable given the automaticity of many expectation effects. For example, the mere statistical dominance of a particular stimulus type is sufficient to bias perceptual interpretations (Chopin & Mamassian 2012). Also, expected words break from continuous flash suppression sooner than unexpected words (Costello et al. 2009).³⁰

All in all, the relation between consciousness on the one hand and categorization, interference, and inference processes on the other hand ranges from non-existent to weak. A much stronger case seems possible for functions like the grouping of image elements according to Gestalt laws and figure-ground segregation. These operations seem to depend strongly on the conscious state, and on conscious perception of the stimuli involved (figure 11). This is surprising, given their relative “simplicity”. For example, the grouping of similarly-oriented or col-linear line segments may be achieved by horizontal connections in the primary visual cortex (Bosking et al. 1997, see above). Figure-ground

segregation—and its neurophysiological correlate—has been successfully modelled in a recurrent network architecture consisting of orientation-selective visual neurons in three hierarchically-organized visual areas, combined with some inhibitory horizontal interactions and excitatory feedback (Roelfsema et al. 2002). Regardless, the experimental data clearly show that if we want to identify visual functions that mark the transition from unconscious processing to conscious vision, grouping according to Gestalt laws (incremental grouping) and figure-ground segregation³¹ (or perceptual organization in general) are our best bets.³²

9 Is it all about distance, or time?

So why do Gestalt grouping and segregation bear such a close relation to consciousness? From a neural perspective, they differ from most other functions in that they depend on interactions between neurons at rather large distances. For example, for a neuron to “know” whether it sits on the figure or the background of the stimulus in figure 5, information has to travel over a distance of about 20 millimetres in the visual cortex.³³ Moreover, the modulations of neural activity that accompany this “knowing” depend on the incremental push-pull interactions between horizontal and feedback connections (Lamme & Roelfsema 2000; Roelfsema et al. 2002; Roelfsema 2006). These require quite extensive processing steps, given that the con-

³¹ Of course one could argue that in the case of a face on a blank background there also is figure-ground segregation. This type of segregation clearly does not depend on consciousness. This touches on the debate on whether categorization is possible without segregation (Wagemans et al. 2012).

³² A promising theory of consciousness holds that conscious representations and states are characterized by the integration of information, or more precisely, on the formation of complexes of integrated information (Tononi 2004, 2008, 2012). That integrated information characterizes consciousness is, however, mainly derived from a set of axioms and introspective or intuitive thought experiments, most of which have already been discussed in the previous text or footnotes (Tononi 2012). What this review of experimental findings however shows is that the “integration of information” comes in many guises, not all of which are equally strongly related to consciousness. A somewhat more precise definition of “integrated information” may be guided by these experimental findings.

³³ The figure is 4 degrees of visual angle wide. Neurons in human V1 with receptive fields at that distance are about 20 mm apart, given a cortical magnification factor of 0.2 degrees per millimetre at 2.0 degrees eccentricity (Duncan & Boynton 2003).

²⁹ In my reading, the predictive coding models are sometimes rather vague about exactly which signal mediates conscious experience. It is often seen to be a combination of the matching process and the posterior, e.g., Seth et al. (2011).

³⁰ But note that this is in fact nothing more than a semantic priming experiment. The results primarily show that if a semantic category has been activated, this category will then break earlier from CFS.

textual Gestalt effects typically manifest themselves at long latency.

Intuitively, seeing an illusion like the Kanizsa triangle, or the contextual shifts in brightness or colour perception discussed above, also seems to depend on “long range” interactions: information travels over large distances in the visual field. But distance travelled over the visual field does not always equal distance travelled in the brain. These phenomena may depend on fairly hardwired and feedforward mechanisms, and their neural correlates typically have relatively short temporal latencies (Von der Heydt et al. 1984). Seemingly, these phenomena tap into mechanisms that have high ecological relevance to the visual system, and are hence solved in a few processing steps, using dedicated feedforward mechanisms. The same holds for all categorization responses in the brain, regardless of their apparent complexity: the progression from low-level to high-level feature detection (including categorization of faces or other complex stimuli) proceeds in a feedforward “sweep” that lasts 100 ms or less (Lamme & Roelfsema 2000).

What emerges is the nagging feeling that consciousness has nothing to do with the seeming complexity or “high-levelness” of a visual function. Whether a visual function depends on consciousness may simply be related to the amount of space that has to be travelled in the brain, how many processing steps have to be taken in between, and hence how much time it takes to complete. This converges onto a thesis that we may call:

The STERP-property of phenomenal representations $=_{\text{DF}}$ conscious representations depend on the spatio-temporally extended neural processing mediated by recurrent interactions.

What that extent is remains to be specified, but has been studied directly by Faivre & Koch (2014), who measured the effects of stimuli made invisible using CFS on the perception of subsequent visible stimuli. Both for apparent motion and for biological motion walkers, it was found that unconscious motion integration only

occurred for relatively short (100 ms) and not for longer (400, 800, 1200ms) temporal intervals. Meng et al. (2007) observed that neural signals representing the spatial filling-in of a grating over a gap in the visual field depended on conscious experience of the grating.³⁴ This suggests that for visual information to literally “bridge a distance” across the visual field, consciousness is required.

The importance of the spatial and temporal extent of neural processing in consciousness also emerges from an entirely different field: that of disorders of consciousness. It is generally believed that there is a gradual decrease of consciousness from the healthy awake state towards minimally conscious, vegetative state and coma. These states also show a gradual decrease in the extent of neural interactions, in both space (Casali et al. 2013) and time (Bekinschtein et al. 2009). Particularly striking is the finding that the presence or absence of consciousness (in this case: the difference between minimally conscious and vegetative state patients) could be classified by simply looking at the amount of “shared symbolic information” in the EEG³⁵ at various distances in the brain. Shared symbolic information at distances of 10 cm and beyond signalled the presence of consciousness, and moreover was indicative of the prognosis of vegetative state patients (whether they would eventually awaken or not). Strikingly, this measure hardly depended on the location of the interactions (King et al. 2013). In other words, whenever and wherever neurons share information at distances of 10 cm or more, there is consciousness.³⁶

Both distance and time are continuous. Arguing that consciousness is related to the temporal or spatial extent of neural processing therefore almost automatically seems to imply that the transition from unconscious to con-

³⁴ It did not depend on attending the grating, however, which is of relevance to the discussion on the relation between attention and consciousness. See below (Lamme 2003, 2004, 2006, 2010a, 2010b).

³⁵ At each electrode, EEG signals were first transformed into symbolic shapes (e.g., up-down-up) for various temporal intervals. Then it was determined to what extent these EEG “symbols” covaried between electrode pairs of various distances, after the exclusion of covariance that was caused by simple volume conduction.

³⁶ Which made me wonder whether any piece of cortex of 10 cm or larger that is held on life support in a petri-dish might have consciousness.

scious processing is gradual rather than discrete. This is not necessarily so, however. Recurrent processing is mediated by highly non-linear interactions, and in such interactions, rather discrete phase transitions are possible (Steyn-Ross et al. 1999; Del Cul et al. 2007; Hwang et al. 2012). It could thus very well be that there is a discrete transition from a phase where information integration is rather limited to a phase that is characterized by extensive information integration, and that this transition depends on the temporal or spatial extent of recurrent interactions.³⁷

Whether the transition from unconscious to conscious processing is discrete or continuous has been argued on different grounds, such as on the distribution of behavioral responses (“seen” versus “not seen”) in relation to manipulations of stimulus variables (Sergent & Dehaene 2004; Overgaard et al. 2006). In signal detection theory, the strength of perceptual information is considered to be continuous, while the decision criterion imposes a discrete boundary between what is reported as “seen” or “not seen”. In its classic form, however, signal detection theory is agnostic about whether consciousness is pre- or post-decisional. Recently, many attempts have been made to incorporate consciousness into the framework of signal detection theory, and in many of these models consciousness is considered post- rather than pre-decisional (Maniscalco & Lau 2012; King & Dehaene 2014)—thus the boundary between the conscious and unconscious is taken to be discrete. Based on neurophysiological findings in the monkey visual cortex, a signal-detection model was devised in which consciousness was considered pre-decisional. In this model, the distribution of sensory information was considered bi-modal, reflecting either a conscious or an unconscious state. The model could explain both the behavioral and neurophysiological findings in the monkey visual cortex, obtained using a variety of stimulus strengths and decision cri-

teria (Supèr et al. 2001). Note that also in this pre-decisional model the conscious–unconscious divide is discrete (or at least bi-modal), rather than gradual.

10 The function of conscious vision

Could it be that Gestalt grouping and figure-ground segregation (of textured images) only happen to go along with consciousness because they take more time; because they require more elaborate computations, not provided by the many dedicated feedforward pathways and modules of the brain? Normally, vision proceeds in a fast and feedforward fashion, where dedicated neurons detect features and categories. Using its hardwired connections, the visual system can swiftly detect the most relevant objects: food, mates, or dangerous animals. Some objects are more difficult to discern, and require prior knowledge or the computation of neighbourhood relations between image elements: food behind a leaf, a sweet versus a sour apple. That takes slightly more—but not too much more—time, because many of the required interactions are hardwired as well. They are hardwired because the visual system has been exposed to these “visual problems” very often, either during evolution or during visual experience. Then there are visual problems that are even more difficult: a camouflaged animal in a crowded forest (figure 7), only visible via subtle differences in overall texture or motion. In this case, all visual resources and mechanisms have to come to the rescue. Only by combining the input from many neurons in a versatile way can the visual “solution” be found. That may be the function of consciousness in the visual domain: to combine the otherwise unconscious modules and mechanisms in a flexible way so as to solve otherwise unresolvable visual problems leading to a second thesis that we may call:

The SUPER-property of phenomenal representations =_{Df} neural representations require consciousness and invoke phenomenality as soon as what needs to be represented can no longer be represented by a single dedicated module or mechanism, yet

³⁷ It could even be that the mere fact that information exchange extends over a particular time and space is critical for that exchange to be accompanied by a conscious sensation. When the same amount of information would be exchanged much slower — as in plants — or much faster — as in a supercomputer — or over a smaller or larger space (as in a microchip or over the internet) no conscious sensation ensues.

requires the interaction of these modules so that a super-positioned representation emerges.

From the point of view of consciousness, a hierarchy of visual functions can then be made. This starts with largely unconscious feature detection and object categorization. These features start to influence each other, and are no longer treated independently, so that categories form that are about the relations between image items (base groupings, short range incremental grouping). With this, there is a transition from the physical properties of the visual input as they are presented to the sensor array to the meaning³⁸ of these properties (e.g., wavelength to colour). During these operations, features and categories are matched with our knowledge and expectations of the world, embedded in the anatomical organization of the visual cortex, aiding in the transformation from visual input towards meaning (inference). Finally, all this information is combined into an organized percept. The longer these operations take, the more distance has to be travelled in the brain, and the more conscious these operations become.³⁹

If nothing interferes, the visual system will always strive towards optimally integrating the available information, so that the richest interpretation of all available information is achieved, and all features have been detected, all inferences have been made, all image elements are combined and all potential ambiguities have been resolved. If this process is cut short, for example by masking or a TMS pulse (Pascual-Leone & Walsh 2001; Silvanto et al. 2005), there is no integrated end-result. And seemingly there is no conscious sensation either. Regardless of this, many features have still been detected, many inferences have been made, and

the brain can use this information to achieve its goals. Behaviour may be influenced, or set into motion (Dehaene et al. 1998). Priming will occur, as well as all sorts of unconscious cognition (Van Gaal & Lamme 2012). Without consciousness, and without maximal integration, the visual system is far from helpless. It can do less, but it can still do a lot.

From this perspective, the function of consciousness in vision is just to enable that last push. That is, to resolve the visual issues that cannot be dealt with otherwise.⁴⁰ And with that, visual functions grow more complex, and evolve from their basic form into more sophisticated versions. A good example comes—once again—from the processing of faces. The core property of face-selective neurons is to respond in a category-selective manner: they distinguish between faces and other objects. They do so from the very first action-potentials that are fired. At that moment, however, category specificity is still very basic, in the sense that all types of faces evoke a similar response (Rolls 1992). At a later moment in time, however, responses typically become more and more specific. In the monkey visual cortex, face cells distinguish between different viewpoints and different emotional expressions of faces with a delay of about 50 milliseconds relative to the categorical face/non-face response (Sugase et al. 1999). View invariant identity representations arise even later, with a delay of about 200 ms (Freiwald & Tsao 2010). At these delays, the face-selective neurons will have established recurrent interactions with lower (and higher) level neurons across the brain, allowing for these more sophisticated classifications to be expressed in the response.

We may thus conclude that face recognition “as we know it”—i.e., not just categorizing face versus non-face, but seeing that face, knowing what it looks like, who it is, and what emo-

³⁸ Note that “meaning” in this context refers to the meaning information has to the organism, shaped by and in accordance with its evolutionary history and ontogenesis (like colour has the “meaning” of the edibility of fruit). It does not refer to “meaning” in any linguistic sense.

³⁹ That may explain why two seemingly similar phenomena like the brightness and colour shifts of figure 2, and the arrival at colour constancy in figure 3 are depend on consciousness in different ways. Colour constancy requires the computation of the full distribution of wavelengths over the entire image, which takes more time than the computations required to compute brightness of adjacent patches.

⁴⁰ Maybe that is the reason why the transition from unconscious to conscious processing also marks the transition between veridical and inferred representations (e.g., from wavelength to colour). Dedicated modules can do their thing in isolation, and therefore have no need to compromise towards a non-veridical representation of the outside world. When modules interact, the necessity may arise to compromise veridical representations to achieve global coherence into the combined super-positioned representation that cannot be represented otherwise.

tion it carries—is a visual function tightly linked to conscious rather than unconscious vision. The main reason for this lies in the fact that in conscious recognition we go beyond simple categorization, and move towards a function where the integration of all possible information about that face (its viewpoint, colour, identity, emotional expression, etc.) is required.

This may raise the question of how we then become conscious of an extremely simple stimulus, such as an oriented black line on a completely white background. With such a simple stimulus, there seems to be no need for any elaborate binding, incremental grouping, or inference. Neurons in the primary visual cortex can detect the line and its orientation within a few action potentials. There seems to be no need to call in the functions that are enabled by conscious processing. So why is it, then, that we still see the black line on the white background?

First, it should be noted that the notion of “simple” stimuli is more complex than one would expect. For example, it was shown that subjects can rapidly detect animals or vehicles in complex natural scenes, even when their attention is simultaneously focused on another task. Discriminating large T’s from L’s, or bisected colour disks from their mirror images was impossible under the same dual task paradigm. Apparently, seemingly simple letter or disk stimuli require more attentive processing than seemingly complex natural scenes (Li et al. 2002), suggesting that they take longer and more elaborate processing. In blindsight, subjects can discriminate lines of different orientations, suggesting that conscious processing is not required for these simple stimuli. However, discrimination performance—although above chance—is typically worse than for consciously-seen line segments, suggesting that something is “missing” from the neural representations formed in blindsight compared to those in conscious vision.

So what might the more elaborate processing steps that lift the unconscious representation of a black line towards a conscious representation of that line be? First, it is known that neurons in many visual areas beyond V1 respond to orientated line segments. At each level,

receptive fields, and hence spatial frequency preferences, differ. This means that (the orientation of) the line segment is represented at many different spatial scales across the visual cortex. Only the integration of these differently-scaled representations, via recurrent interactions, yields a precise and conscious representation. The same holds for other properties of the “simple” line segment, such as its colour, its depth, and its relation to the background.⁴¹ Indeed, oriented lines are fairly easy to mask (in fact easier than faces), indicating that their conscious percept depends on more elaborate processing steps than expected for such a simple stimulus.

11 The impact of conscious vision on the brain

If a particular visual problem has to be dealt with often, the brain will start to build connections so that the problem can be resolved more rapidly. Visual problems that require long and elaborate processing will eventually be resolved in milliseconds. By building new and dedicated connections, elaborate processing steps may be simplified into a fast and short set of interactions. Conscious processing will turn into unconscious processing, because conscious processing has triggered perceptual learning that in turn evokes synaptic changes that create new “dedicated modules” that can do the job unconsciously. This leads to a third thesis:⁴²

The LEARN-property of phenomenal representations =_{Df} neural representations that require consciousness and invoke phenomenality, at the same time evoke synaptic plasticity mechanisms and learning,

⁴¹ Even something as simple as a white background will give the black line another visual “meaning” than a yellow background, a green background, or a textured background. The same point has been formulated by Giulio Tononi (2004, 2008, 2012): a conscious representation is conscious because it differentiates from the endless other potential representations that could have been. In this case: the oriented black line on the white background is one of the endless possible configurations of lines on backgrounds, and only by integrating the information of line and background is it known which of these configurations is actually present.

⁴² Similar ideas exist in the context of motor learning: a task that first requires extensive conscious practice will gradually become more and more automatic, up to the point where it can be executed fully unconsciously.

in an attempt to make these representations less dependent on consciousness and invoking less phenomenality.

Indeed, there are several arguments for linking consciousness to perceptual learning. Plasticity in the visual cortex comes in many temporal and spatial scales. There are fast- and short-range adaptations or recalibrations, expressed in altered stimulus-response dependencies (e.g., contrast normalization). But receptive fields may also change in size or feature selectivity when exposed to repeated stimulation. Receptive fields literally grow or shift position when their surrounds are stimulated but the receptive field is not (Gilbert & Wiesel 1992). Prolonged depletion of input leads to the induction of new connectivity via fast axonal sprouting of horizontal connections (Yamahachi et al. 2009). Horizontal connections in particular play an important role in both immediate and longer term plasticity of the visual cortex (Gilbert et al. 1996). The repeated execution of Gestalt grouping via the same connections may therefore induce learning (Gilbert et al. 2001), as, for example, is observed in the learning of texture segregation (Karni & Sagi 1991) or in the gradual improvement of contour integration during childhood development (Kovács et al. 1999). In addition, perceptual learning induces a reorganization of the areas involved in encoding the learned object—a process that is mediated by feedback connections (Sigman & Gilbert 2000; Sigman et al. 2005). It seems that the neural machinery that mediates Gestalt grouping and segregation is also the machinery that mediates perceptual learning.

Furthermore, feedback and horizontal connections have been linked to the molecular mechanisms of neural plasticity. A key component in neural plasticity is the NMDA receptor pathway, and in the monkey, NMDA receptor blocking using APV reduces contextual figure-ground modulation (Self et al. 2012). Similarly, in humans, figure-ground segregation is impaired using Ketamine, an anaesthetic which selectively blocks the NMDA receptor at low doses (Meuwese et al. 2013). Also, it was found that Ketamine at sub-anaesthetic doses inter-

feres with the learning of Mooney figures. Mooney figures are high-contrast versions of images that are hard to recognize when you don't know what the image is about. Once you have seen its original natural contrast version, however, the Mooney image is readily recognizable. It was found that the neural representation of Mooney images starts to resemble that of their natural versions once they are learned. Ketamine disrupts this rapid learning process, but only in V1, and not in higher visual areas, indicating that feedback from higher areas to V1 is selectively disrupted by Ketamine (Van Loon et al. submitted).

In sum, there are strong indications that link conscious visual processing and its neural machinery—horizontal and feedback connection—are linked to perceptual learning and the molecular mechanisms involved. This may open up a path to a more molecular understanding of consciousness. In addition, it provides us with a clear idea about the function of consciousness: that of building a new repertoire of visual functions, so that eventually conscious processing is no longer necessary.

It must be noted however, that the link between consciousness and learning is controversial. Many instances of “unconscious” perceptual learning exist (e.g., Gutnisky et al. 2009; Seitz et al. 2009; Seitz & Watanabe 2003; Schwiedrzik et al. 2011). An important issue here, however, is whether these are cases of learning without conscious experience of the stimuli that induce the learning, or whether they are instances of learning without cognitive access or attention to these stimuli (see Meuwese et al. 2013). A further clarification of the role of consciousness in learning is required.

12 The dolphins of consciousness research

I have examined the defining characteristics of conscious versus unconscious vision. Incremental grouping and segregation according to Gestalt laws seems to be a defining characteristic of conscious vision. Other visual phenomena and functions, like interference or inference, are less

strongly linked. Feature detection and higher-level categorization clearly do not mark the transition from unconscious to conscious vision. From a neural perspective, it can be argued that conscious processing is linked to those operations that require spatially and temporally extended processing, where neurons engage in incremental interactions involving many steps. These processes are selectively dependent on horizontal and feedback connections. Moreover, these interactions induce learning, as they operate along highly plastic neural pathways, and use the molecular machinery that is directly involved in neural plasticity.

We can now start using these defining characteristics to answer more difficult questions. Is there consciousness in the right half-brain of a split brain patient (Sperry 1984)? Is there consciousness without attention (Koch & Tsuchiya 2012)? Is there consciousness in neglect or extinction (Lamme 2003)? Is it appropriate to talk about inattentional “blindness”, where people do not remember having seen something while their attention was engaged elsewhere? What exactly happens during change blindness (Simons & Rensink 2005)? Is there consciousness in animals (Edelman & Seth 2009), or in a vegetative state (Owen et al. 2006)? These are the “dolphins” of consciousness research, situations that are hard to position in the current taxonomy of conscious versus unconscious, because much controversy exists about the presence or absence of conscious experience in those conditions. With this, I hope to have given some usable arguments that can settle such controversies. My claim would simply be that whenever we see the defining properties of conscious vision that have been laid out here (i.e., incremental Gestalt grouping and segregation), there is conscious vision, regardless of whether there is conscious access or report (e.g., Scholte et al. 2006). More in general, the more fruitful stance towards consciousness would be to let all the available evidence converge into general theses, such as those derived here, and then take these as the defining characteristics of conscious processing and consciousness, regardless of whether they fit our introspective intuition of

what consciousness is or should be. Defining consciousness as the process that builds on spatio-temporally extended neural processing (**STERP property**), that enables the building of super-positioned representations that individual modules cannot provide (**SUPER property**), and that evokes synaptic plasticity and learning (**LEARN property**) yields clear defining characteristics. These characteristics go a great length towards elucidating important features of phenomenality (its integrated nature, Gestalt properties), towards explaining the nature of conscious experience (perceptual organization, interference, inference), and are hinting towards a potential function of consciousness (learning) and its molecular basis. What I consider irrelevant characteristics (such as the ability to report about an experience, see Lamme 2010a, 2010b) generally do no such explaining. It is better to build a taxonomy of conscious versus unconscious processing on defining characteristics than on irrelevant ones. That has helped a lot in positioning dolphins in the taxonomy of species. It will also help a lot in positioning the wild amalgam of phenomena that the field of consciousness research has produced so far. And it will enable us to give consciousness its proper ontological status. But I have already contributed to that discussion extensively elsewhere (Lamme 2003, 2004, 2006, 2010a, 2010b), so I will lay that to rest here.

At the crack of dawn, something magical happens. Night turns into day, life springs, vibrations fill the air. We know, it is just the earth rotating. But a very fundamental transition it remains. Unconscious or conscious processing, it's all neurons doing their job, firing action potentials, exchanging chemicals, transferring information. But somehow, suddenly, they “turn on the light”. You see. You have a conscious sensation of that dawn. Isn't it beautiful? You should take a picture of it.

Acknowledgements

This work was supported by an ERC Advanced Investigator Grant (DEFCON1, nr 230355) to Victor Lamme.

References

- Ahmed, B., Cordery, P. M., McLelland, D., Bair, W. & Krug, K. (2012). Long-range clustered connections within extrastriate visual area V5/MT of the rhesus macaque. *Cerebral Cortex*, 22 (1), 60-73. [10.1093/cercor/bhr072](https://doi.org/10.1093/cercor/bhr072)
- Albert, M. K. & Hoffman, D. D. (2000). The generic-viewpoint assumption and illusory contours. *Perception*, 29 (3), 303-312.
- Alkire, M. T., Hudetz, A. G. & Tononi, G. (2008). Consciousness and anesthesia. *Science*, 322 (5903), 876-80. [10.1126/science.1149213](https://doi.org/10.1126/science.1149213)
- Allman, J., Miezin, F. & McGuinness, E. (1985). Stimulus specific responses from beyond the classical receptive field: Neurophysiological mechanisms for local-global comparisons in visual neurons. *Annual Review of Neuroscience*, 8, 407-30. [10.1146/annurev.ne.08.030185.002203](https://doi.org/10.1146/annurev.ne.08.030185.002203)
- Almeida, J., Mahon, B. Z., Nakayama, K. & Caramazza, A. (2008). Unconscious processing dissociates along categorical lines. *Proceedings of the National Academy of Sciences of the United States of America*, 105 (39), 15214-15218. [10.1073/pnas.0805867105](https://doi.org/10.1073/pnas.0805867105)
- Barbur, J. L., de Cunha, D., Williams, C. B. & Plant, G. (2004). Study of instantaneous color constancy mechanisms in human vision. *Journal of Electronic Imaging*, 13 (1), 15-28. [10.1117/1.1636491](https://doi.org/10.1117/1.1636491)
- Barbur, J. L. & Spang, K. (2008). Colour constancy and conscious perception of changes of illuminant. *Neuropsychologia*, 46 (3), 853-863. [10.1016/j.neuropsychologia.2007.11.032](https://doi.org/10.1016/j.neuropsychologia.2007.11.032)
- Bekinschtein, T. A., Dehaene, S., Rohaut, B., Tadel, F., Cohen, L. & Naccache, L. (2009). Neural signature of the conscious processing of auditory regularities. *Proceedings of the National Academy of Sciences*, 106 (5), 1672-1677. [10.1073/pnas.0809667106](https://doi.org/10.1073/pnas.0809667106)
- Block, N. (1996). How can we find the neural correlate of consciousness? *Trends in Neurosciences*, 19 (11), 456-459. [10.1016/S0166-2236\(96\)20049-9](https://doi.org/10.1016/S0166-2236(96)20049-9)
- (2005). Two neural correlates of consciousness. *Trends in Cognitive Sciences*, 9 (2), 46-52. [10.1016/j.tics.2004.12.006](https://doi.org/10.1016/j.tics.2004.12.006)
- (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences*, 30 (5-6), 481-499. [10.1017/S0140525X07002786](https://doi.org/10.1017/S0140525X07002786)
- Bosking, W. H., Zhang, Y., Schofield, B. & Fitzpatrick, D. (1997). Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. *The Journal of Neuroscience*, 17 (6), 2112-2127.
- Boyer, J. L., Harrison, S. & Ro, T. (2005). Unconscious processing of orientation and color without primary visual cortex. *Proceedings of the National Academy of Sciences*, 102 (46), 16875-16879. [10.1073/pnas.0505332102](https://doi.org/10.1073/pnas.0505332102)
- Breitmeyer, B. G., Ogmen, H. G., Ro, T. & Singhal, N. S. (2004). Unconscious color priming occurs at stimulus-not percept-dependent levels of processing. *Psychological Science*, 15 (3), 198-202. [10.1111/j.0956-7976.2004.01503009.x](https://doi.org/10.1111/j.0956-7976.2004.01503009.x)
- Breitmeyer, B. G., Ogmen, H. G., Ro, T., Ogmen, H. & Todd, S. (2007). Unconscious, stimulus-dependent priming and conscious, percept-dependent priming with chromatic stimuli. *Perception & Psychophysics*, 69 (4), 550-557. [10.3758/BF03193912](https://doi.org/10.3758/BF03193912)
- Breitmeyer, B. G. & Ogmen, H. (2000). Recent models and findings in visual backward masking: A comparison, review, and update. *Perception & Psychophysics*, 62 (8), 1572-1595. [10.3758/BF03212157](https://doi.org/10.3758/BF03212157)
- Bullier, J., Hupé, J. M., James, A. C. & Girard, P. (2001). The role of feedback connections in shaping the responses of visual cortical neurons. *Progress in Brain Research*, 134, 193-204.
- Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., Casarotto, S., Bruno, M. A., Laureys, S., Tononi, G. & Massimini, M. (2013). A theoretically based index of consciousness independent of sensory processing and behavior. *Science Translational Medicine*, 5 (198), 198ra105. [10.1126/scitranslmed.3006294](https://doi.org/10.1126/scitranslmed.3006294)
- Chalmers D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2 (3), 200-219.
- Chopin, A. & Mamassian, P. (2012). Predictive properties of visual adaptation. *Current Biology*, 22 (7), 622-626. [10.1016/j.cub.2012.02.021](https://doi.org/10.1016/j.cub.2012.02.021)
- Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy*, 78 (2), 67-90. [10.2307/2025900](https://doi.org/10.2307/2025900)
- Cohen, M. A. & Dennett, D. C. (2011). Consciousness cannot be separated from function. *Trends in Cognitive Sciences*, 15 (8), 358-364. [10.1016/j.tics.2011.06.008](https://doi.org/10.1016/j.tics.2011.06.008)
- Costello, P., Jiang, Y., Baartman, B., McGlennen, K. & He, S. (2009). Semantic and subword priming during binocular suppression. *Consciousness and Cognition*, 18 (2), 375-382. [10.1016/j.concog.2009.02.003](https://doi.org/10.1016/j.concog.2009.02.003)
- Cowey, A. & Heywood, C. A. (1997). Cerebral achromatopsia: Colour blindness despite wavelength processing. *Trends in Cognitive Sciences*, 1 (4), 133-139. [10.1016/S1364-6613\(97\)01043-7](https://doi.org/10.1016/S1364-6613(97)01043-7)

- Cox, M. A., Schmid, M. C., Peters, A. J., Saunders, R. C., Leopold, D. A. & Maier, A. (2013). Receptive field focus of visual area V4 neurons determines responses to illusory surfaces. *Proceedings of the National Academy of Sciences*, 110 (42), 17095-17100. [10.1073/pnas.1310806110](https://doi.org/10.1073/pnas.1310806110)
- Crick, F. & Koch, C. (1998). Consciousness and neuroscience. *Cerebral Cortex*, 8 (2), 97-107. <http://www.ncbi.nlm.nih.gov/pubmed/9542889>
- (2003). A framework for consciousness. *Nature Neuroscience*, 6 (2), 119-126. [10.1038/nn0203-119](https://doi.org/10.1038/nn0203-119)
- Cumming, B. G. & Parker, A. J. (1997). Responses of primary visual cortical neurons to binocular disparity without depth perception. *Nature*, 389 (6648), 280-283. [10.1038/38487](https://doi.org/10.1038/38487)
- Dehaene, S., Naccache, L., Le Clec'H, G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., Van de Moortele, P. F. & Le Bihan, D. (1998). Imaging unconscious semantic priming. *Nature*, 395 (6702), 597-600. [10.1038/26967](https://doi.org/10.1038/26967). <http://www.ncbi.nlm.nih.gov/pubmed/9783584>
- Dehaene, S., Jobert, A., Naccache, L., Ciuciu, P., Poline, J. B., Le Bihan, D. & Cohen, L. (2004). Letter binding and invariant recognition of masked words: Behavioral and neuroimaging evidence. *Psychological Science*, 15 (5), 307-313. [10.1111/j.0956-7976.2004.00674.x](https://doi.org/10.1111/j.0956-7976.2004.00674.x)
- Dehaene, S., Changeux, J. P., Naccache, L., Sackur, J. & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: A testable taxonomy. *Trends in Cognitive Sciences*, 10 (5), 204-211. [10.1016/j.tics.2006.03.007](https://doi.org/10.1016/j.tics.2006.03.007)
- Dehaene, S. & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, 79 (1-2), 1-37. <http://www.ncbi.nlm.nih.gov/pubmed/11164022>
- Del Cul, A., Baillet, S. & Dehaene, S. (2007). Brain dynamics underlying the nonlinear threshold for access to consciousness. *PLoS Biology*, 5 (10), e260. [10.1371/journal.pbio.0050260](https://doi.org/10.1371/journal.pbio.0050260)
- Dennett, D. C. (1993). *Consciousness explained*. New York, NY: Penguin. (pp. 889-892). [10.2307/2108259](https://doi.org/10.2307/2108259)
- Dow, B. M., Snyder, A. Z., Vautin, R. G. & Bauer, R. (1981). Magnification factor and receptive field size in foveal striate cortex of the monkey. *Experimental Brain Research*, 44 (2), 213-228. <http://www.ncbi.nlm.nih.gov/pubmed/7286109>
- Duncan, R. O. & Boynton, G. M. (2003). Cortical magnification within human primary visual cortex correlates with acuity thresholds. *Neuron*, 38 (4), 659-671. [10.1016/S0896-6273\(03\)00265-4](https://doi.org/10.1016/S0896-6273(03)00265-4)
- Edelman, D. B. & Seth, A. K. (2009). Animal consciousness: A synthetic approach. *Trends in Neurosciences*, 32 (9), 476-484. [10.1016/j.tins.2009.05.008](https://doi.org/10.1016/j.tins.2009.05.008)
- Enns, J. T. & Di Lollo V. (2000). What's new in visual masking? *Trends in Cognitive Sciences*, 4 (9), 345-352. [10.1016/S1364-6613\(00\)01520-5](https://doi.org/10.1016/S1364-6613(00)01520-5)
- Fahrenfort, J. J., Scholte, H. S. & Lamme, V. A. F. (2007). Masking disrupts reentrant processing in human visual cortex. *Journal of Cognitive Neuroscience*, 19 (9), 1488-1497. [10.1162/jocn.2007.19.9.1488](https://doi.org/10.1162/jocn.2007.19.9.1488)
- Fahrenfort, J. J., Snijders, T. M., Heinen, K., Van Gaal, S., Scholte, H. S. & Lamme, V. A. F. (2012). Neuronal integration in visual cortex elevates face category tuning to conscious face perception. *Proceedings of the National Academy of Sciences*, 109 (52), 21504-21509. [10.1073/pnas.1207414110](https://doi.org/10.1073/pnas.1207414110)
- Fahrenfort, J. J. & Lamme, V. A. F. (2012). A true science of consciousness explains phenomenology: Comment on Cohen and Dennett. *Trends in Cognitive Sciences*, 16 (3), 138-140. [10.1016/j.tics.2012.01.004](https://doi.org/10.1016/j.tics.2012.01.004)
- Faivre, N. & Koch, C. (2014). Temporal structure coding with and without awareness. *Cognition*, 131 (3), 404-414. [10.1016/j.cognition.2014.02.008](https://doi.org/10.1016/j.cognition.2014.02.008)
- Fang, F. & He, S. (2005). Cortical responses to invisible objects in the human dorsal and ventral pathways. *Nature Neuroscience*, 8 (10), 1380-1385. [10.1038/nn1537](https://doi.org/10.1038/nn1537)
- Foster, K. H., Gaska, J. P., Nagler, M. & Pollen, D. A. (1985). Spatial and temporal frequency selectivity of neurones in visual cortical areas V1 and V2 of the macaque monkey. *Journal of Physiology*, 365, 331-363.
- Freiwald, W. A. & Tsao, D. Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, 330 (6005), 845-851. [10.1126/science.1194908](https://doi.org/10.1126/science.1194908)
- Gazzaniga, M. S. (2005). Forty-five years of split-brain research and still going strong. *Nature Reviews Neuroscience*, 6 (8), 653-659. [10.1038/nrn1723](https://doi.org/10.1038/nrn1723)
- Gilbert, C. D., Das, A., Ito, M., Kapadia, M. & Westheimer, G. (1996). Spatial integration and cortical dynamics. *Proceedings of the National Academy of Sciences*, 93 (2), 615-622.
- Gilbert, C. D., Sigman, M. & Crist, R. E. (2001). The neural basis of perceptual learning. *Neuron*, 31 (5), 681-97. <http://www.ncbi.nlm.nih.gov/pubmed/11567610>
- Gilbert, C. D. & Wiesel, T. N. (1989). Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex. *The Journal of Neuroscience*, 9 (7), 2432-42. <http://www.ncbi.nlm.nih.gov/pubmed/2746337>

- (1992). Receptive field dynamics in adult primary visual cortex. *Nature*, 356 (6365), 150-2. [10.1038/356150a0](https://doi.org/10.1038/356150a0). <http://www.ncbi.nlm.nih.gov/pubmed/1545866>
- Girard, P., Hupé J. M., James A. C. & Bullier, J. (2001). Feedforward and feedback connections between areas V1 and V2 of the monkey have similar rapid conduction velocities. *Journal of Neurophysiology*, 85 (3), 1328-1331. <http://www.ncbi.nlm.nih.gov/pubmed/11248002>
- Gobbini, M. I., Gors, J. D., Halchenko, Y. O., Hughes, H. C. & Cipolli, C. (2013). Processing of invisible social cues. *Consciousness and Cognition*, 22 (3), 765-770. [10.1016/j.concog.2013.05.002](https://doi.org/10.1016/j.concog.2013.05.002)
- Goodale, M. A. & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, 15 (1), 20-25. [10.1016/0166-2236\(92\)90344-8](https://doi.org/10.1016/0166-2236(92)90344-8)
- Gutnisky, D. A., Hansen, B. J., Iliescu, B. F. & Dragoi, V. (2009). Attention alters visual plasticity during exposure-based learning. *Current Biology*, 19 (7), 555-560. [10.1016/j.cub.2009.01.063](https://doi.org/10.1016/j.cub.2009.01.063)
- Harris, J. J., Schwarzkopf, D. S., Song, C., Bahrami, B. & Rees, G. (2011). Contextual illusions reveal the limit of unconscious visual processing. *Psychological Science*, 22 (3), 399-405. [10.1177/0956797611399293](https://doi.org/10.1177/0956797611399293)
- He, S., Cavanagh, P. & Intriligator, J. (1996). Attentional resolution and the locus of visual awareness. *Nature*, 383 (6598), 334-337. [10.1038/383334a0](https://doi.org/10.1038/383334a0)
- Hess, R. F., Beaudot, W. H. & Mullen, K. T. (2001). Dynamics of contour integration. *Vision Research*, 41 (8), 1023-1037. [10.1016/S0042-6989\(01\)00020-7](https://doi.org/10.1016/S0042-6989(01)00020-7)
- Hubel, D. H. (1982). Exploration of the primary visual cortex, 1955-78. *Nature*, 299 (5883), 515-524. <http://www.ncbi.nlm.nih.gov/pubmed/6750409>
- Hubel, D. H. & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, 195 (1), 215-243.
- Hupé, J. M., James, A. C., Payne, B. R., Lomber, S. G., Girard, P. & Bullier, J. (1998). Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. *Nature*, 394 (6695), 784-787. [10.1038/29537](https://doi.org/10.1038/29537)
- Hupé, J. M., James, A. C., Girard, P. & Bullier, J. (2001). Response modulations by static texture surround in area V1 of the macaque monkey do not depend on feedback connections from V2. *Journal of Neurophysiology*, 85 (1), 146-63.
- Hupé, J. M., James, A. C., Girard, P., Lomber, S. G., Payne, B. R. & Bullier, J. (2001). Feedback connections act on the early part of the responses in monkey visual cortex. *Journal of Neurophysiology*, 85 (1), 134-145.
- Hwang, E., Kim, S., Han, K. & Choi, J. H. (2012). Characterization of phase transition in the thalamocortical system during anesthesia-induced loss of consciousness. *PLoS One*, 7 (12), e50580. [10.1371/journal.pone.0050580](https://doi.org/10.1371/journal.pone.0050580)
- Jiang, Y., Costello, P., Fang, F., Huang, M. & He, S. (2006). A gender- and sexual orientation-dependent spatial attentional effect of invisible images. *Proceedings of the National Academy of Sciences*, 103 (45), 17048-17052. [10.1073/pnas.0605678103](https://doi.org/10.1073/pnas.0605678103)
- Kamermans, M., Kraaij, D. A. & Spekreijse, H. (1998). The cone/horizontal cell network: A possible site for color constancy. *Visual Neuroscience*, 15 (5), 787-797.
- Kapadia, M. K., Ito, M., Gilbert, C. D. & Westheimer, G. (1995). Improvement in visual sensitivity by changes in local context: Parallel studies in human observers and in V1 of alert monkeys. *Neuron*, 15 (4), 843-856. [10.1016/0896-6273\(95\)90175-2](https://doi.org/10.1016/0896-6273(95)90175-2)
- Karni, A. & Sagi, D. (1991). Where practice makes perfect in texture discrimination: Evidence for primary visual cortex plasticity. *Proceedings of the National Academy of Sciences of the USA*, 11 (88), 4966-4970. [10.1016/j.cub.2009.01.063](https://doi.org/10.1016/j.cub.2009.01.063)
- Kim, C. Y. & Blake, R. (2005). Psychophysical magic: Rendering the visible 'invisible'. *Trends in Cognitive Sciences*, 9 (8), 381-388. [10.1016/j.tics.2005.06.012](https://doi.org/10.1016/j.tics.2005.06.012)
- King, J. R. & Dehaene, S. (2014). A model of subjective report and objective discrimination as categorical decisions in a vast representational space. *Philosophical Transactions of the Royal Society B*, 369 (1641), 1471-2970. [10.1098/rstb.2013.0204](https://doi.org/10.1098/rstb.2013.0204)
- King, J. R., Sitt, J. D., Faugeras, F., Rohaut, B., El Karoui, I., Cohen, L., Naccache, L. & Dehaene, S. (2013). Information sharing in the brain indexes consciousness in noncommunicative patients. *Current Biology*, 23 (19), 1914-1919. [10.1016/j.cub.2013.07.075](https://doi.org/10.1016/j.cub.2013.07.075). <http://www.ncbi.nlm.nih.gov/pubmed/24076243>
- Knierim, J. J. & Van Essen, D. C. (1992). Neuronal responses to static texture patterns in area V1 of the alert macaque monkey. *Journal of Neurophysiology*, 67 (4), 961-980.
- Koch, C. & Tsuchiya, N. (2012). Attention and consciousness: Related yet different. *Trends in Cognitive Sciences*, 16 (2), 103-105. [10.1016/j.tics.2011.11.012](https://doi.org/10.1016/j.tics.2011.11.012)
- Kolb, F. C. & Braun, J. (1995). Blindsight in normal observers. *Nature*, 377 (6547), 336-338. [10.1038/377336a0](https://doi.org/10.1038/377336a0)
- Kouider, S., Eger, E., Dolan, R. & Henson, R. N. (2009). Activity in face-responsive brain regions is modulated by invisible, attended faces: Evidence from masked priming. *Cerebral Cortex*, 19 (1), 13-23. [10.1093/cercor/bhn048](https://doi.org/10.1093/cercor/bhn048)

- Kovács, I., Kozma, P., Fehér, A. & Benedek, G. (1999). Late maturation of visual spatial integration in humans. *Proceedings of the National Academy of Sciences*, 96 (21), 12204-12209. [10.1073/pnas.96.21.12204](https://doi.org/10.1073/pnas.96.21.12204)
- Lamme, V. A. F. (1995). The neurophysiology of figure-ground segregation in primary visual cortex. *Journal of Neuroscience*, 15 (2), 1605-1615.
- (2003). Why visual attention and awareness are different. *Trends in Cognitive Sciences*, 7 (1), 12-18. [10.1016/S1364-6613\(02\)00013-X](https://doi.org/10.1016/S1364-6613(02)00013-X)
- (2004). Separate neural definitions of visual consciousness and visual attention; a case for phenomenal awareness. *Neural Networks*, 17 (5-6), 861-872. [10.1016/j.neunet.2004.02.005](https://doi.org/10.1016/j.neunet.2004.02.005)
- (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*, 10 (11), 494-501. [10.1016/j.tics.2006.09.001](https://doi.org/10.1016/j.tics.2006.09.001)
- (2010a). How neuroscience will change our view on consciousness. *Cognitive Neuroscience*, 1 (3), 204-220. [10.1080/17588921003731586](https://doi.org/10.1080/17588921003731586)
- (2010b). What introspection has to offer, and where its limits lie. *Cognitive Neuroscience*, 1 (3), 232-240. [10.1080/17588928.2010.502224](https://doi.org/10.1080/17588928.2010.502224)
- Lamme, V. A. F., Van Dijk, B. W. & Spekreijse, H. (1993). Contour from motion processing occurs in primary visual cortex. *Nature*, 363 (6429), 541-543. [10.1038/363541a0](https://doi.org/10.1038/363541a0)
- Lamme, V. A. F., Supér, H. & Spekreijse, H. (1998a). Feedforward, horizontal, and feedback processing in the visual cortex. *Current Opinion in Neurobiology*, 8 (4), 529-35. <http://www.ncbi.nlm.nih.gov/pubmed/9751656>
- Lamme, V. A. F., Zipser, K. & Spekreijse, H. (1998b). Figure-ground activity in primary visual cortex is suppressed by anesthesia. *Proceedings of the National Academy of Sciences of the United States of America*, 95 (6), 3263-8. <http://www.ncbi.nlm.nih.gov/pubmed/9501251>
- Lamme, V. A. F., Rodriguez-Rodriguez, V. & Spekreijse, H. (1999). Separate processing dynamics for texture elements, boundaries and surfaces in primary visual cortex of the macaque monkey. *Cerebral Cortex*, 9 (4), 406-13. <http://www.ncbi.nlm.nih.gov/pubmed/10426419>
- Lamme, V. A. F., Supér, H., Landman, R., Roelfsema, P. R. & Spekreijse, H. (2000). The role of primary visual cortex (V1) in visual awareness. *Vision Research*, 40 (10-12), 1507-21. <http://www.ncbi.nlm.nih.gov/pubmed/10788655>
- Lamme, V. A. F., Zipser, K. & Spekreijse, H. (2002). Masking interrupts figure-ground signals in V1. *Journal of Cognitive Neuroscience*, 14 (7), 1044-53. [10.1162/089892902320474490](https://doi.org/10.1162/089892902320474490). <http://www.ncbi.nlm.nih.gov/pubmed/12419127>
- Lamme, V. A. F. & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, 23 (11), 571-9. <http://www.ncbi.nlm.nih.gov/pubmed/11074267>
- Lamme, V. A. F. & Spekreijse, H. (2000). Modulations of primary visual cortex activity representing attentive and conscious scene perception. *Frontiers in Bioscience*, 5, D232-43. <http://www.ncbi.nlm.nih.gov/pubmed/10704153>
- Li, F. F., Van Rullen, R., Koch, C. & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences*, 99 (14), 9596-9601. [10.1073/pnas.092277599](https://doi.org/10.1073/pnas.092277599)
- Macknik, S. L. & Livingstone, M. S. (1998). Neuronal correlates of visibility and invisibility in the primate visual system. *Nature Neuroscience*, 1 (2), 144-149. [10.1038/393](https://doi.org/10.1038/393)
- Maier J., Dagnelie G., Sprekreijse H. & Van Dijk B. W. (1987). Principal components-analysis for source localization of VEPs in man. *Vision Research*, 27 (2), 165-177. [10.1016/0042-6989\(87\)90179-9](https://doi.org/10.1016/0042-6989(87)90179-9)
- Malach, R., Amir, Y., Harel, M. & Grinvald, A. (1993). Relationship between intrinsic connections and functional architecture revealed by optical imaging and in vivo targeted biocytin injections in primate striate cortex. *Proceedings of the National Academy of Sciences*, 90 (22), 10469-10473.
- Maniscalco, B. & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, 21 (1), 422-430. [10.1016/j.concog.2011.09.021](https://doi.org/10.1016/j.concog.2011.09.021)
- Marcel, A. J. (1998). Blindsight and shape perception: Deficit of visual consciousness or of visual function? *Brain*, 121 (8), 1565-1588. [10.1093/brain/121.8.1565](https://doi.org/10.1093/brain/121.8.1565)
- Meng, M., Ferneyhough, E. & Tong, F. (2007). Dynamics of perceptual filling-in of visual phantoms revealed by binocular rivalry. *Journal of Vision*, 7 (13). [10.1167/7.13.8](https://doi.org/10.1167/7.13.8)
- Meuwese, J. D., Post, R. A., Scholte, H. S. & Lamme, V. A. F. (2013). Does perceptual learning require consciousness or attention? *Journal of Cognitive Neuroscience*, 25 (10), 1579-1596. [10.1162/jocn_a_00424](https://doi.org/10.1162/jocn_a_00424)
- Meuwese, J. D., Scholte, H. S. & Lamme, V. A. F. (2014). Latent memory of unattended stimuli reactivated by practice: An fMRI study on the role of consciousness and attention in learning. *PLoS One*, 9 (3), e90098. [10.1371/journal.pone.0090098](https://doi.org/10.1371/journal.pone.0090098). <http://www.ncbi.nlm.nih.gov/pubmed/24603676>

- Moore, T., Rodman, H. R., Repp, A. B. & Gross, C. G. (1995). Localization of visual stimuli after striate cortex damage in monkeys: Parallels with human blindsight. *Proceedings of the National Academy of Sciences*, 92 (18), 8215-8218. <http://www.ncbi.nlm.nih.gov/pubmed/7667270>
- Moradi, F., Koch, C. & Shimojo, S. (2005). Face adaptation depends on seeing the face. *Neuron*, 45 (1), 169-175. [10.1016/j.neuron.2004.12.018](https://doi.org/10.1016/j.neuron.2004.12.018)
- Moutoussis, K. & Zeki, S. (2002). The relationship between cortical activation and perception investigated with invisible stimuli. *Proceedings of the National Academy of Sciences*, 99 (14), 9527-9532. [10.1073/pnas.142305699](https://doi.org/10.1073/pnas.142305699)
- Mudrik, L., Breska, A., Lamy, D. & Deouell, L. Y. (2011). Integration without awareness: Expanding the limits of unconscious processing. *Psychological Science*, 22 (6), 764-770. [10.1177/0956797611408736](https://doi.org/10.1177/0956797611408736)
- Nakayama, K. & Shimojo, S. (1992). Experiencing and perceiving visual surfaces. *Science*, 257 (5075), 1357-1363.
- Nothdurft, H. C., Gallant, J. L. & Van Essen, D. C. (1999). Response modulation by texture surround in primate area V1: Correlates of “popout” under anesthesia. *Visual Neuroscience*, 16 (1), 15-34.
- Oram, M. W. & Perrett, D. I. (1992). Time course of neural responses discriminating different views of the face and head. *Visual Neuroscience*, 68 (1), 70-84.
- Overgaard, M., Rote, J., Mouridsen, K. & Ramsøy, T. Z. (2006). Is conscious perception gradual or dichotomous? A comparison of report methodologies during a visual task. *Consciousness and Cognition*, 15 (4), 700-708. [10.1016/j.concog.2006.04.002](https://doi.org/10.1016/j.concog.2006.04.002)
- Owen, A. M., Coleman, M. R., Boly, M., Davis, M. H., Laureys, S. & Pickard, J. D. (2006). Detecting awareness in the vegetative state. *Science*, 313 (5792), 1402. [10.1126/science.1130197](https://doi.org/10.1126/science.1130197)
- Panichello, M. F., Cheung, O. S. & Bar, M. (2012). Predictive feedback and conscious visual experience. *Frontiers in Psychology*, 3. [10.3389/fpsyg.2012.00620](https://doi.org/10.3389/fpsyg.2012.00620)
- Pascual-Leone, A. & Walsh, V. (2001). Fast backprojections from the motion to the primary visual area necessary for visual awareness. *Science*, 292 (5516), 510-512. [10.1126/science.1057099](https://doi.org/10.1126/science.1057099)
- Rao, R. P. & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2 (1), 79-87. [10.1038/4580](https://doi.org/10.1038/4580)
- Robichaud, L. & Stelmach, L. B. (2003). Inducing blindsight in normal observers. *Psychonomic Bulletin & Review*, 10 (1), 206-209.
- Rock, I. & Palmer, S. (1990). The legacy of Gestalt psychology. *Scientific American*, 263 (6), 84-90. <http://www.ncbi.nlm.nih.gov/pubmed/2270461>
- Roe, A. W., Lu, H. D. & Hung, C. P. (2005). Cortical processing of a brightness illusion. *Proceedings of the National Academy of Sciences*, 102 (10), 3869-3874. [10.1073/pnas.0500097102](https://doi.org/10.1073/pnas.0500097102)
- Roelfsema, P. R. (2006). Cortical algorithms for perceptual grouping. *Annual Review of Neuroscience*, 29, 203-227. [10.1146/annurev.neuro.29.051605.112939](https://doi.org/10.1146/annurev.neuro.29.051605.112939)
- Roelfsema, P. R., Lamme, V. A. F., Spekreijse, H. & Bosch, H. (2002). Figure-ground segregation in a recurrent network architecture. *Journal of Cognitive Neuroscience*, 14 (4), 525-537. [10.1162/08989290260045756](https://doi.org/10.1162/08989290260045756)
- Rolls, E. T. (1992). Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas. *Philosophical Transactions of the Royal Society*, 335 (1273), 11-21. [10.1098/rstb.1992.0002](https://doi.org/10.1098/rstb.1992.0002)
- Rolls, E. T. & Tovee, M. J. (1994). Processing speed in the cerebral cortex and the neurophysiology of visual masking. *Proceedings of the Royal Society B: Biological Sciences*, 257 (1348), 9-15. [10.1098/rspb.1994.0087](https://doi.org/10.1098/rspb.1994.0087). <http://www.ncbi.nlm.nih.gov/pubmed/8090795>
- Rossi, A. F., Rittenhouse, C. D. & Paradiso, M. A. (1996). The representation of brightness in primary visual cortex. *Science*, 273 (5278), 1104-1107.
- Rossi, A. F. & Paradiso, M. A. (1999). Neural correlates of perceived brightness in the retina, lateral geniculate nucleus, and striate cortex. *Journal of Neuroscience*, 19 (14), 6145-56.
- Scholte, H. S., Witteveen, S. C., Spekreijse, H. & Lamme, V. A. F. (2006). The influence of inattention on the neural correlates of scene segmentation. *Brain Research*, 1076 (1), 106-115. [10.1016/j.brainres.2005.10.051](https://doi.org/10.1016/j.brainres.2005.10.051)
- Schwiedrzik, C. M., Singer, W. & Melloni, L. (2011). Subjective and objective learning effects dissociate in space and in time. *Proceedings of the National Academy of Sciences*, 108 (11), 4506-4511. [10.1073/pnas.1009147108](https://doi.org/10.1073/pnas.1009147108)
- Seghier, M. L. & Vuilleumier, P. (2006). Functional neuroimaging findings on the human perception of illusory contours. *Neuroscience & Biobehavioral Reviews*, 30 (5), 595-612. [10.1016/j.neubiorev.2005.11.002](https://doi.org/10.1016/j.neubiorev.2005.11.002)
- Seitz, A. R., Kim, D. & Watanabe, T. (2009). Rewards evoke learning of unconsciously processed visual stimuli in adult humans. *Neuron*, 61 (5), 700-707. [10.1016/j.neuron.2009.01.016](https://doi.org/10.1016/j.neuron.2009.01.016)

- Seitz, A. R. & Watanabe, T. (2003). Psychophysics: Is subliminal learning really passive? *Nature*, 422 (6927), 36. [10.1038/422036a](https://doi.org/10.1038/422036a)
- Self, M. W., Kooijmans, R. N., Supèr, H., Lamme, V. A. F. & Roelfsema, P. R. (2012). Different glutamate receptors convey feedforward and recurrent processing in macaque V1. *Proceedings of the National Academy of Sciences of the United States of America*, 109 (27), 11031-6. [10.1073/pnas.1119527109](https://doi.org/10.1073/pnas.1119527109).
<http://www.ncbi.nlm.nih.gov/pubmed/22615394>
- Self, M. W., Van Kerkoerle, T., Supèr, H. & Roelfsema, P. R. (2013). Distinct roles of the cortical layers of area V1 in figure-ground segregation. *Current Biology*, 23 (21), 2121-2129. [10.1016/j.cub.2013.09.013](https://doi.org/10.1016/j.cub.2013.09.013)
- Sergent, C. & Dehaene, S. (2004). Is consciousness a gradual phenomenon? Evidence for an all-or-none bifurcation during the attentional blink. *Psychological Science*, 15 (11), 720-728.
[10.1111/j.0956-7976.2004.00748.x](https://doi.org/10.1111/j.0956-7976.2004.00748.x).
<http://www.ncbi.nlm.nih.gov/pubmed/15482443>
- Seth, A. K. (2010). The grand challenge of consciousness. *Frontiers in Psychology*, 1. [10.3389/fpsyg.2010.00005](https://doi.org/10.3389/fpsyg.2010.00005)
- Seth, A. K., Suzuki, K. & Critchley, H. D. (2011). An interoceptive predictive coding model of conscious presence. *Frontiers in Psychology*, 2, 395. [10.3389/fpsyg.2011.00395](https://doi.org/10.3389/fpsyg.2011.00395).
<http://www.ncbi.nlm.nih.gov/pubmed/22291673>
- Shapley, R. & Hawken, M. J. (2011). Color in the cortex: Single- and double-opponent cells. *Vision Research*, 51 (7), 701-717. [10.1016/j.visres.2011.02.012](https://doi.org/10.1016/j.visres.2011.02.012)
- Sigman, M., Pan, H., Yang, Y., Stern, E., Silbersweig, D. & Gilbert, C. D. (2005). Top-down reorganization of activity in the visual pathway after learning a shape identification task. *Neuron*, 46 (5), 823-835.
[10.1016/j.neuron.2005.05.014](https://doi.org/10.1016/j.neuron.2005.05.014)
- Sigman, M. & Gilbert, C. D. (2000). Learning to find a shape. *Nature Neuroscience*, 3 (3), 264-269.
[10.1038/72979](https://doi.org/10.1038/72979)
- Silvanto, J., Lavie, N. & Walsh, V. (2005). Double dissociation of V1 and V5/MT activity in visual awareness. *Cerebral Cortex*, 15 (11), 1736-1741. [10.1093/cercor/bhi050](https://doi.org/10.1093/cercor/bhi050)
- Simons, D. J. & Rensink, R. A. (2005). Change blindness: past, present, and future. *Trends in Cognitive Sciences*, 9 (1), 16-20. [10.1016/j.tics.2004.11.006](https://doi.org/10.1016/j.tics.2004.11.006).
<http://www.ncbi.nlm.nih.gov/pubmed/15639436>
- Snodderly, D. M. & Gur, M. (1995). Organization of striate cortex of alert, trained monkeys (Macaca fascicularis): Ongoing activity, stimulus selectivity, and widths of receptive field activating regions. *Journal of Neurophysiology*, 74 (5), 2100-2125.
- Sperry, R. (1984). Consciousness, personal identity and the divided brain. *Neuropsychologia*, 22 (6), 661-673.
[10.1016/0028-3932\(84\)90093-9](https://doi.org/10.1016/0028-3932(84)90093-9)
- Stein, T., Hebart, M. N. & Sterzer, P. (2011). Breaking continuous flash suppression: A new measure of unconscious processing during interocular suppression? *Frontiers in Human Neuroscience*, 5. [10.3389/fnhum.2011.00167](https://doi.org/10.3389/fnhum.2011.00167)
- Steyn-Ross, M. L., Steyn-Ross, D. A., Sleight, J. W. and Liley, D. T. (1999). Theoretical electroencephalogram stationary spectrum for a white-noise-driven cortex: Evidence for a general anesthetic-induced phase transition. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 60 (6 Pt B), 7299-7311. [10.1103/PhysRevE.60.7299](https://doi.org/10.1103/PhysRevE.60.7299)
- Stoerig, P. & Cowey, A. (1989). Wavelength sensitivity in blindsight. *Nature*, 342 (6252), 916-918. [10.1038/342916a0](https://doi.org/10.1038/342916a0)
- Straube, T., Mothes-Lasch, M. & Miltner, W. H. (2011). Neural mechanisms of the automatic processing of emotional information from faces and voices. *British Journal of Social Psychology*, 102 (4), 830-848.
[10.1111/j.2044-8295.2011.02056.x](https://doi.org/10.1111/j.2044-8295.2011.02056.x)
- Sugase, Y., Yamane, S., Ueno, S. & Kawano, K. (1999). Global and fine information coded by single neurons in the temporal visual cortex. *Nature*, 400 (6747), 869-873. [10.1038/23703](https://doi.org/10.1038/23703)
- Sugihara, T., Qiu, F. T. & Von der Heydt, R. (2011). The speed of context integration in the visual cortex. *Journal of Neurophysiology*, 106 (1), 374-385. [10.1152/jn.00928.2010](https://doi.org/10.1152/jn.00928.2010)
- Supèr, H., Spekreijse, H. & Lamme, V. A. F. (2001). Two distinct modes of sensory processing observed in monkey primary visual cortex (V1). *Nature Neuroscience*, 4 (3), 304-310. [10.1038/85170](https://doi.org/10.1038/85170)
- Supèr, H., Van der Togt, C., Spekreijse, H. & Lamme, V. A. F. (2003). Internal state of monkey primary visual cortex (V1) predicts figure-ground perception. *Journal of Neuroscience*, 23 (8), 3407-3414.
- Supèr, H. & Lamme, V. A. F. (2007). Altered figure-ground perception in monkeys with an extra-striate lesion. *Neuropsychologia*, 45 (14), 3329-3334.
[10.1016/j.neuropsychologia.2007.07.001](https://doi.org/10.1016/j.neuropsychologia.2007.07.001)
- Swets, J. A., Green, D. M., Getty, D. J. & Swets, J. B. (1978). Signal detection and identification at successive stages of observation. *Perception & Psychophysics*, 23 (4), 275-289.
- Sáry, G., Köteles, K., Kaposvári, P., Lenti, L., Csifcsák, G., Frankó, E., Benedek, G. & Tompa, T. (2008). The representation of Kanizsa illusory contours in the monkey inferior temporal cortex. *European Journal of Neuroscience*, 28 (10), 2137-2146. [10.1111/j.1460-9568.2008.06499.x](https://doi.org/10.1111/j.1460-9568.2008.06499.x)

- Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, 19, 109-139. [10.1146/annurev.ne.19.030196.000545](https://doi.org/10.1146/annurev.ne.19.030196.000545). <http://www.ncbi.nlm.nih.gov/pubmed/8833438>
- Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, 5 (42). [10.1186/1471-2202-5-42](https://doi.org/10.1186/1471-2202-5-42). <http://www.ncbi.nlm.nih.gov/pubmed/15522121>
- (2008). Consciousness as integrated information: A provisional manifesto. *Biological Bulletin*, 215 (3), 216-242.
- (2012). Integrated information theory of consciousness: an updated account. *Archives Italiennes de Biologie*, 150 (4), 293-329.
- Tononi, G. & Massimini, M. (2008). Why does consciousness fade in early sleep? *Annals of the New York Academy of Sciences*, 1129, 330-334. [10.1196/annals.1417.024](https://doi.org/10.1196/annals.1417.024)
- Tsuchiya, N. & Koch, C. (2005). Continuous flash suppression reduces negative afterimages. *Nature Neuroscience*, 8 (8), 1096-1101. [10.1038/nm1500](https://doi.org/10.1038/nm1500)
- Van der Togt, C., Kalitzin, S., Spekreijse, H., Lamme, V. A. F. & Supèr, H. (2006). Synchrony dynamics in monkey V1 predict success in visual detection. *Cerebral Cortex*, 16 (1), 136-148. [10.1093/cercor/bhi093](https://doi.org/10.1093/cercor/bhi093)
- Van Gaal, S. & Lamme, V. A. F. (2012). Unconscious high-level information processing: Implication for neurobiological theories of consciousness. *Neuroscientist*, 18 (3), 287-301. [10.1177/1073858411404079](https://doi.org/10.1177/1073858411404079)
- Van Loon, A. M., Fahrenfort, J. J., Van der Velde, B., Lirk, P. B., Vulink, N. C. C., Hollmann, M. W., Scholte, H. S. and Lamme, V. A. F. (submitted). NMDA receptor antagonist ketamine distorts object recognition by reducing feedback to early visual cortex. *The Journal of Neuroscience*
- Von der Heydt, R., Peterhans, E. & Baumgartner, G. (1984). Illusory contours and cortical neuron responses. *Science*, 224 (4654), 1260-1262.
- Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M. and Von der Heydt, R. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization. *Psychological Bulletin*, 138 (6), 1172-1217. [10.1037/a0029333](https://doi.org/10.1037/a0029333)
- Wang, L., Weng, X. & He, S. (2012). Perceptual grouping without awareness: Superiority of Kanizsa triangle in breaking interocular suppression. *PLoS One*, 7 (6), e40106. [10.1371/journal.pone.0040106](https://doi.org/10.1371/journal.pone.0040106)
- Weiskrantz, L. (1996). Blindsight revisited. *Current Opinion in Neurobiology*, 6 (2), 215-220. [10.1016/S0959-4388\(96\)80075-4](https://doi.org/10.1016/S0959-4388(96)80075-4)
- Wokke, M. E., Sligte, I. G., Scholte, H. S. & Lamme, V. A. F. (2012). Two critical periods in early visual cortex during figure-ground segregation. *Brain and Behavior*, 2 (6), 763-777. [10.1002/brb3.91](https://doi.org/10.1002/brb3.91)
- Yamahachi, H., Marik, S. A., McManus, J. N., Denk, W. & Gilbert, C. D. (2009). Rapid axonal sprouting and pruning accompany functional reorganization in primary visual cortex. *Neuron*, 64 (5), 719-729. [10.1016/j.neuron.2009.11.026](https://doi.org/10.1016/j.neuron.2009.11.026)
- Zipser, K., Lamme, V. A. F. & Schiller, P. H. (1996). Contextual modulation in primary visual cortex. *Journal of Neuroscience*, 16 (22), 7376-7389.

Consciousness as Inference in Time

A Commentary on Victor Lamme

Lucia Melloni

Unraveling the neural correlates of conscious remains one of the great challenges of our time. Victor Lamme proposes that neural integration through feedback loops is what differentiates conscious from unconscious processing. Here, I review his hypothesis, focusing on the spatial scale of integration as well as the possible neural mechanisms involved. I go on to show that any theory of the neural correlates of consciousness is incomplete if it cannot account for how prior knowledge shapes perception and how this form of integration occurs. Finally, I propose that integration across moments in time is a crucial but hitherto neglected aspect of conscious perception, which creates the “flow” of conscious experience.

Keywords

Active sensing | Expectations | Flow of consciousness | Neural correlates of consciousness | Predictive coding

Commentator

Lucia Melloni

lucia.melloni@brain.mpg.de

Max Planck Institute for Brain Research

Frankfurt a. M., Germany

Target Author

Victor Lamme

victorlamme@gmail.com

Universiteit van Amsterdam

Amsterdam, Netherlands,

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität

Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University

Melbourne, Australia

1 Qualia 2.1: Integration is key but is it all?

Why do we see the way we *see*? How is our perception different from the way a photograph is acquired on the sensor chip of a digital camera? It seems obvious that we do not see an image made of individual pixels but an integrated, smooth, colourful, and vivid image. What is the neural substrate of this marvellous capacity that makes us feel and experience the way we do? These are the central questions that Victor Lamme sets out to address in his paper *The Crack of Dawn: Perceptual Functions and Neural Mechanisms that Mark the Transition from Unconscious Processing to Conscious Vision*.

This is by no means an easy task, even when one stays away from the difficult problem of qualia or “what it is like to be” (Nagel 1974). The question of how awareness arises has preoccupied philosophers and scientists for centuries, and while significant progress has been made in recent decades we are still far from reaching a conclusion (Dehaene 2014; Koch 2004). One thing is clear however: success in understanding the neural machinery that instantiates consciousness rests on identifying the fundamental features that characterise a state as conscious and that distinguish it from unconscious states.

A remarkable discovery of the past century is that a significant portion of all mental operations, including fairly complex ones such as decision-making and perceptual categorisation, can be carried out unconsciously. Take the case of language: while it seems effortless to understand the words that you are currently reading, you do not have conscious access to the syntactic processes that ultimately allow you to grasp the relations between the elements of this sentence and thus its meaning. These complex mental operations occur “behind the scene” of consciousness. Given that so many intricate processes can operate unconsciously, one cannot but wonder what consciousness is good for. Which mental processes *require* consciousness, if any? And if so, what really distinguishes conscious from unconscious cognition? Victor Lamme offers a stimulating and comprehensive review of processes in vision that can be performed outside the realm of awareness. The list is long and may be surprising (also see Kouider & Dehaene 2007), ranging from detection of simple (e.g., oriented lines) and complex features (e.g., faces; Almeida et al. 2013; de Gardelle et al. 2011; Del Zotto et al. 2013), to mathematical operations such as abstract comparisons between quantities (Greenwald et al. 2003), to triggering of motor plans (Dehaene et al. 1998), and even error-related responses to stimuli that fully escape our consciousness (Cohen et al. 2009).

What do we need consciousness for, then? Lamme proposes that consciousness is required when all sources of information need to be integrated. For instance, when we see a face, we can not only *detect* that is a face, a process that can be performed unconsciously, but also *identify* it as that of our friend Billy, whom we have not seen in ten years and that we *remember* warmly from our childhood. Consciousness brings this *unified* moment in which all comes together: previous experiences are retrieved from memory (e.g., do we have reason to like Billy?) and unified with the context of the current experience (e.g., where are we now?), but also intertwined with predictions for future actions (e.g., would we like to engage in a conversation?). Thus, in one single moment, past,

present and future come together and form a unified conscious experience. Many scientists nowadays agree that conscious experience provides an abstract summary of all available sources of information, from which many features are filtered out and reinterpreted in a format that is most useful for further actions, thoughts, deliberations, and chain operations that cannot be processed by non-conscious processors (Lamme this collection; Baars 2002; Dehaene 2014; Melloni & Singer 2010). Hence, what reaches our perception is a highly processed, “interpreted” version of the world. One key intuition is that the unification and “interpretation” of the experience that reaches our consciousness is achieved through the activation of myriads of neurons that signal individual features, but that it is by virtue of *integrating* their information through dynamic interactions (for example via synchronous coordination of their activity or via feedback processes) that a coherent experience across senses, space, and time comes about.

An important caveat is that integration of information *per se* is unlikely to distinguish conscious from unconscious processing as integration of many features can also proceed unconsciously (Dehaene et al. 1998; Gaillard et al. 2009; Lin & He 2009; Melloni et al. 2007; Melloni & Rodriguez 2007; Mudrik et al. 2014). In fact, integration through convergence is a key principle of the wiring of the brain, which explains the mere existence of feature-selective neurons that respond to motion, shape, or complex stimuli such as faces, and that process information in an unconscious manner. If it is not integration *per se*, then what kind of integration are we talking about?¹ We and others (Melloni & Singer 2010; Thompson & Varela 2001;

¹ Giulio Tononi (2004; Tononi & Koch 2008) argues that not only integration but also differentiation/segregation (e.g., distinguishing a particular state from all possible other states) is characteristic of conscious states. However, even when both conditions are met, say integration through convergence is observed in FFA and differentiated from other states, e.g., there is no activation in PPA, an area selective to processing places, and thus there is no guarantee that this would constitute a conscious state. In fact, experimental evidence suggests that such feature-selective processing can indeed proceed unconsciously, for example in the case of face processing under conditions of masking (de Gardelle et al. 2011), continuous flash suppression (Almeida et al. 2013), and in blindsight patients (Del Zotto et al. 2013).

Varela et al. 2001) have previously argued for a distinction between *local* and *global* integration, and proposed that the spatial scale of integration differentiates between unconscious and conscious states: unconscious processing is observed when local integration occurs within the divergent-convergent feedforward architecture; conscious processing however requires long-range integration through neural synchronization, which integrates information across the various levels of the cortical processing hierarchy.

Indeed, in recent years, a wealth of experimental studies (Aru et al. 2012; Gaillard et al. 2009; Hipp et al. 2011; Melloni et al. 2007; Melloni & Rodriguez 2007) have provided support to the idea that long-range integration through synchronous coupling is a mechanism for conscious perception, and that the spatial scale of synchronisation strongly correlates with the perceptual outcome. For example, we have shown that masked words are only consciously perceived when accompanied by a burst of long-distance synchronization in the gamma band, while unconscious processing, even up to a semantic level, elicits only local gamma oscillations (Melloni et al. 2007; Melloni & Rodriguez 2007). Although controversy still persists as to whether long-range integration necessitates the involvement of particular brain areas (Dehaene 2014; Edelman & Tononi 2000) or not (Lamme this collection; Melloni & Singer 2010), it is reassuring to witness some convergence on the results that have even led to clinical applications (e.g., coma classification, King et al. 2013). In his most recent work, Victor Lamme now also assigns a central role to the spatial scale of the integration for consciousness, joining an ever-increasing number of researchers proposing long-range integration as key to consciousness (Dehaene & Changeux 2011; Edelman & Tononi 2000; Melloni & Singer 2010; Thompson & Varela 2001). An interesting point of divergence from other theories is that while Lamme assigns a particular role to feedback and horizontal connections in the integration of information for consciousness, other theories, including our own, hypothesise that it is the synchronisation of neural populations that glues all experiences into one, thereby instantiating con-

sciousness. As empirical data and theoretical considerations continue to accumulate, we expect that this and other pressing challenges such as identifying how far is “long” in the brain, or whether “long” involves the activation of specific neural cell populations, specific areas, and/or a specified number of nodes will become addressable.

However, imagine those questions have been addressed and we know that integration on a particular spatial scale is key to consciousness; would we have understood what consciousness is or how it comes about? Here I propose that we would not, as any theory that does not account for two fundamental, hitherto neglected aspects of conscious experience will fall short of explaining consciousness. In particular, our experience is never an island in isolation, but instead is shaped by previous knowledge, by priors that stem from the preceding context or from our history of learning. These priors determine our perception; and thus understanding how they become integrated is paramount to explaining consciousness. However, an even more pressing problem is that conscious experience unfolds over time, whereby the recent past moulds the current moment, which in turn creates predictions for moments to come, i.e., the future. How all those temporal processes intertwine and define our experience (the flow of consciousness) is something that most research has neglected. In the following sections I will review current research that we and others have undertaken with the purpose of raising awareness of these overlooked integrative properties of conscious experience and the challenges that they entail for the study of consciousness.

2 Consciousness as an inferential process and the consequences for the neural mechanism of conscious perception

One central and characteristic feature of conscious perception is its constructive nature. In contrast to unconscious cognition, which is directly driven by sensory stimulation, the images that reach consciousness often bear little resemblance to reality. Indeed, percepts in our

mind can be understood as useful distortions of reality in which only specific parts of the physical input are represented while being enriched with a model of the world that has been learned and that provides context to the current moment. In the words of [Heinz von Foerster \(1984\)](#), “the world, as we perceive it, is our own invention”. To provide a striking example of this, consider the image on the right (Figure 1) and try to figure out what it shows. Most people at first see a collection of black and white blobs, much like the input that strikes our retina—a raw, uninterpreted signal. Now, rotate the page upside down. Voila! You will clearly see a face (do you recognize whose face it is?). Remarkably, you can turn the page back and you will continue seeing the face. Once you have recognized the image, the visual system has created a *prior*, an expectation that enriches perception. This example is not mere curiosity. Most of our behaviour and perceptions are based on predictions: we do not wait for visual input to impinge our eyes, we *actively* look for it. We cannot, however, initiate a rational search for an object without making *predictions* about “what” it is, “where” it is likely to be, and even “when” it is likely to be there. The brain’s ability to make predictions and to mould its data gathering accordingly is thus essential for its ability to evaluate options, make life-critical decisions, and generate adaptive behaviour.

While the constructive nature of perception is undeniable and may even appear as one of its defining features, surprisingly little research has been carried out to understand how previous experience interacts with consciousness. Most importantly, the scientific community has not embraced an understanding of consciousness in the context of a flow of experience in which every moment is integrated with past moments and interfaced with expectations about what will happen in the future (but see [Varela 1999](#)). A possible reason for neglecting the contribution of previous experience is that this integration of past with present moments has been understood as a process of “unconscious inference” (following [von Helmholtz 1866/1962](#)), or, in Victor Lamme’s words, in the context of the “automaticity of the many ex-

pectation effects.” However, this inferential process is carried out in the backstage of consciousness, and it is only the result that we consciously experience. This bears resemblance to syntactic analysis, which is also carried out automatically and unconsciously, but is paramount to conscious access to meaning. Without unconscious syntactic analysis we would not be able to “consciously” understand text; nor is its automatic activation under our control. In the same vein, our conscious perception would be totally different if prior knowledge did not help us enrich or even construct our experience, endowing it with meaning. In fact, it has been proposed that alterations in perception, i.e., the defragmented sensory experience observed in schizophrenics and autistic people can be the result of a deficit in this inferential process ([Jardri & Deneve 2013](#); [Pellicano & Burr 2012](#)), underscoring the fundamental role that perceptual inference plays in conscious perception.

One promising framework within which the influence of previous experience through unconscious inference can be understood is the Bayesian framework. When applied to perception, each mathematically-formulated ingredient of this framework can be assigned a perceptual counterpart, with previous experience referring to the prior, the current moment referring to the likelihood, unconscious inference referring to Bayes rule (which combines the prior with the likelihood in an optimal way), and the result—our perception—referring to (the peak of) the posterior distribution. This idea has recently proven to be a powerful tool for understanding perception not only in terms of modelling behaviour, but also as a theoretical framework for understanding how perception arises in the brain. A prominent implementation of the latter is Predictive Coding ([Friston 2010](#)). This theory postulates that the brain builds models (priors) of the world based on previous experience, which are used to explain the current inputs. This occurs iteratively *across all levels in the cortical hierarchy* with the goal of minimising predictions errors, i.e., the difference between what is expected and the incoming sensory input, which are energetically costly. This minimization process can either be achieved by chan-

ging the way the system samples its environment, or by changing its models. Relevant for this discussion is the idea that perceptual inference, in the Predictive Coding framework, implies that all levels in the hierarchy reach an agreement, i.e., minimise all prediction errors, much like the idea of a unified/integrative moment as proposed by Victor Lamme and others (Dehaene 2014; Edelman & Tononi 2000; Melloni & Singer 2010). While Predictive Coding by itself is currently agnostic as to whether such unified agreement represents a conscious state, the central tenet that integration across all levels is what the system strives for still holds. This allows for the formulation of interesting, testable predictions about the Neural Correlates of Consciousness (NCC).

In recent years research in my lab has focused on understanding how previous experience enriches perception, how expectations alter the NCC, and how this can be understood within the Predictive Coding framework. The central idea that motivated these studies was to test whether or not the NCC are context independent, i.e., impervious to the influence of expectations, as many theories implicitly postulate. To test this hypothesis we presented subjects with illusory letters, that is letters whose borders were not explicitly defined but instead required the activation of figure-ground segregation cues. We reasoned that providing subjects with a prior, i.e., knowing which letter would be presented next, would facilitate the figure-ground segregation process, making an initially invisible letter clearly visible. In line with our expectations, we observed that the threshold of conscious perception is not fixed but instead changes depending on the availability of previous knowledge: subjects are able to perceive a stimulus on the basis of minimal sensory information when they have a clear expectation. We were able to confirm this result in a series of different paradigms in which expectations could be generated online from recent experience as in the example of the letter given above (Melloni et al. 2011; Schwiedrzik et al. 2014), drawn from memory based on prior exposure to clearly visible natural images (Aru et al. 2012), stem from a life-long history of association between letters

and colour as in grapheme-colour synaesthesia (van Leeuwen et al. 2013), or result from systematic training as in perceptual learning (Schwiedrzik et al. 2009, 2011). These studies allowed us to test not only whether the behavioural threshold of conscious perception is fixed, but also how previous knowledge would affect the neural “construction” of conscious percepts.



Figure 1: Can you recognize what this is? If not, rotate the image. Note that once you turn it back around the object is now clear.

A first hypothesis we derived from the Predictive Coding framework was that the presence of strong priors should have an effect of how quickly content reaches awareness. If conscious perception is the result of a process that iterates until information is consistent between the different levels of the hierarchy (Di Lollo et al. 2000), i.e., until all prediction errors are minimised, then having a better model of the input based on prior knowledge may speed up this process. Indeed and contrary to the com-

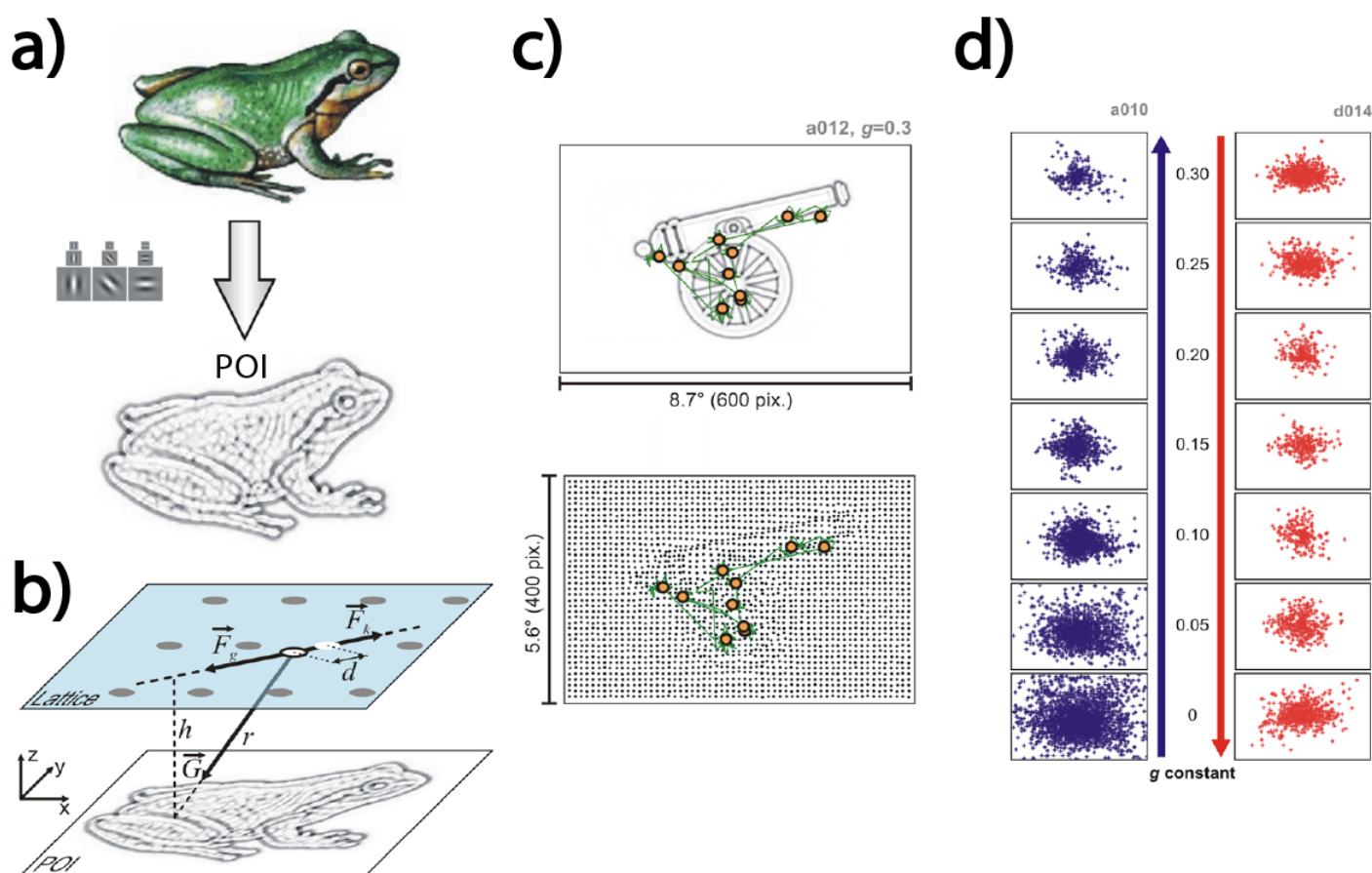


Figure 2: (a) Original images are filtered through a series of gabor wavelets, which allows the estimation of the points of maximal local information (Points of Maximal Information, POI) in the source image. (b) Dots of an elastic lattice are created by mapping the POI in the projection plane, and attracting them by the projection F_0 of a gravitational force G . (c) Pattern of saccades/fixations when subjects recognise a stimulus and its underlying POI map. (d) Pattern of fixations for stimuli of different degradation levels from high degradation (0) to low degradation (0.30). Dots in blue correspond to fixations when subjects do not have an expectation of the stimuli, dots in red correspond to patterns of fixation observed in the presence of expectations. Note that in the presence of expectations, the distribution of fixations are much less scattered. From [Moca et al. \(2011\)](#).

mon belief that information processing in the brain has a fixed latency, we observed that the NCC shifts in time when a prior is available. While the electrophysiological difference between seen and unseen letters occurred around 300ms when it exclusively depended on sensory evidence, it occurred as early as 200ms when priors were available ([Melloni et al. 2011](#)). Thus, priors sped up information processing by 100ms. These results have important implications for the search for the NCC as they show that conscious processing is not bound to a particular time, but can flexibly adjust its timing depending on the task at hand, the readiness of the system, or the presence of expectations.

They also pose a challenge to theories that postulate that the NCC always occur late, as proposed by [Victor Lamme \(this collection\)](#) or [Stanislas Dehaene \(2014\)](#).

A second prediction that follows from the principle of minimising prediction errors is that in the presence of priors, activity in lower areas can be “explained away” by priors in higher brain areas ([Murray et al. 2004](#)); this entails that when inputs can be fully predicted based on previous experience, they do not elicit prediction errors. To test this hypothesis, we took the same study to the MEG and performed source localisation. Here, we found that priors sparsify the networks involved in processing the

stimulus, such that when a prior is present only the brain areas that are most diagnostic to the stimulus features are activated (Mayer et al. in preparation). All alternative interpretations of the stimulus are thus “explained away”. Thus, consciousness and its neural correlates appear as mobile targets, which adjust their locus in the presence of expectations. This poses a further challenge to the search for the NCC, as not only the timing, but also the location of neural activation does not appear as a diagnostic feature for the NCC.

Finally, Predictive Coding also suggests that priors may be used to change the way information is sampled, as the models derived from previous experience can be used to optimise the search for the most relevant information (Friston et al. 2012). Only rarely do we keep our gaze still and wait for the world to bring novel information; instead, we scan images through rhythmic patterns of eye movements accompanied by fixations. This active sensing view implies that perception is not a passive phenomenon in which the system waits for information to hit the sensory transducers, but instead an active process that seeks information through exploratory routines (Melloni et al. 2009; Schroeder et al. 2010). To test whether and how priors affect the sampling of information we developed stimuli for which we could quantify the local information content at each point (Figure 2) and determined the efficiency of information extraction based on eye movements in the presence or absence of expectations. Figure 2 shows that when subjects have prior knowledge of the object they are trying to perceive, they can immediately orient their eyes to areas of most diagnostic information for the perception of an object. At the same time, the sampling of information becomes sparser, concentrating eye movements to maximally informative areas (Moca et al. 2011). This implies that priors direct our exploratory motor routines, thus optimising perception.

Overall, these studies show that previous experience enriches the contents of consciousness and fundamentally changes the way information is processed in our brain, enhancing speed and efficiency. This raises questions for

theories that propose a fixed latency or neural locus for conscious access, but also complicates the quest for the NCC, as they turn out to differ in time and location depending on the precision and accuracy of expectations. Although current formulations of Predictive Coding do not make specific predictions about consciousness, this framework may nevertheless prove to be an important starting point in trying to understand these effects. In fact, more explicit theoretical links between Predictive Coding and consciousness are now being worked out (e.g., Clark 2013; Hohwy 2013; Seth et al. 2011)—after all, Predictive Coding has been framed as a unifying theory of the brain (Friston 2010), which would fall short if consciousness was left unexplained.

3 The neglected dimension of consciousness: Time and the flow of consciousness

But is that all? One dimension of our experience that is often neglected is time. Of course, time is an implicit component of previous experience, however, it may also be revealing to consider time by itself. In fact, living organisms seldom encounter a static image in isolation, but are instead confronted with a flow of temporally-correlated sensory inputs (Schwartz et al. 2007). Imagine for instance a tennis match, and picture the tennis ball flying over the field. If queried, you could easily estimate where the ball is, but also where it was a second ago and where it will be in a few milliseconds. Event-objects of the conscious mind² thus per definition unfold in time and we also act in time: we make use of current and previous input to figure out the most appropriate response predicting their consequences. There is thus a continuum of interdependencies along the time dimension whereby every past moment is *integrated* with the present and projected into the future, giving rise to the flow of consciousness. The same way

² We are usually conscious of objects, and become so by virtue of their being differentiated from the background, but also because their internal features are linked or bound in some way. Objects and their internal features do not need to be static entities but can have temporal dynamics, i.e., they develop or change in time. In this case, they become events (and thus event-objects of the conscious mind).

we have been thinking about the integration of multiple source of information *within* a given moment of time, such as multiple features of a single object, there is thus integration *across* time. A case in point is strikingly vivid perceptual aftereffects, such as the waterfall illusion, where viewing motion in one direction for several seconds causes a subsequently presented static image to move in the opposite direction (Purkinje 1820). Such effects are not limited to basic perceptual features such as motion direction, colour, or orientation, but also affect high-level percepts such as the perceived gender of faces (Webster et al. 2004), numerosity (Burr & Ross 2008), or gaze direction (Jenkins et al. 2006); and they are not limited to fleeting illusions that vanish almost instantaneously, but may persist for days or even weeks (Jones & Holding 1975). This indicates that our current experience is embedded into a continuous flow of previous experience at multiple time scales, ranging from lifelong experience with our environment to short-term, moment-by-moment effects that arise from our most recent encounters, even if just milliseconds ago.

The past thus leaves traces (predictions) that determine the current contents of consciousness. This has the consequence that the contents of consciousness represent an aggregate of imprints from the past and the present moment that jointly promote a sense of *stability over time*. However, through which mechanism these interdependencies affect our perception is currently unclear. Experimentally, the multiple time-scales of previous experience are particularly evident when subjects are confronted with sequences of multistable stimuli such as the Necker cube.³ Because the sensory information these stimuli provide by themselves is insufficient to determine perception, they are particularly susceptible to the effects of previous experience. Under these conditions, one can observe two different effects that temporal dependencies entail: on the one hand, an *attractive* effect, which increases the likelihood of continuing to perceive the same stimulus, and on the other hand a *repulsive* effect, which increases the like-

lihood of perceiving something different. The former is often referred to as hysteresis, priming, stabilisation, or perceptual memory, while the latter is commonly known as perceptual adaptation.

Recently, Chopin & Mamassian (2012) studied the temporal dynamics of these serial dependencies, addressing the question of which part of the perceptual history the system retains and how remote and recent experiences differentially determine perception. They observed a remarkable dissociation between long stretches of time that occurred in the remote past (in their case several minutes) and short stretches of time that had just recently occurred (a few seconds ago): while the former had a positive correlation with perception, and thus ensured stability over time (hysteresis), the latter had a negative correlation to perception, that is, it promoted alternative interpretations (adaptation). These two timescales indicate that previous experience can act along at least two separate timescales and hence, that there may be several mechanisms at work. Using functional magnetic resonance imaging, we set out to further elucidate how these effects are implemented in the brain, how the brain entertains these two opposing processes without mutual interference, and what determines their direction (Schwiedrzik et al. 2014). Presenting multistable visual stimuli sequentially, we found that although affecting our perception concurrently, hysteresis and adaptation map into distinct cortical networks: a widespread network of higher-order visual and fronto-parietal areas was involved in hysteresis, while adaptation was confined to early visual areas (areas V2/V3). Importantly, hysteresis and adaptation bear a differential relation with whether or not the stimuli were consciously perceived: while adaptation was present even if the adapting interpretation was not consciously perceived (in agreement with previous reports, e.g., Hock et al. 1996), hysteresis depended on what was previously consciously perceived. Hence, conscious experiences in the past affected the present experience, preserving continuity in time, while unconscious processing had the opposite effect, bringing change and novelty to perception.

³ But they are by no means limited to ambiguous stimuli (Fischer & Whitney 2014; Treisman 1984).

This brings us back to the question of neural integration, indicating that even in the case of *integration over time*, the spatial scale at which neuronal processing occurs determines whether content enters awareness or not: in the case of hysteresis, a conscious moment is integrated in time with another conscious moment, which involves a widespread cortical network, while in the case of adaptation, prior information is only integrated within a local module, which happens irrespective of whether this prior information is consciously experienced or not, similar to Lamme's "base grouping". This interpretation fits with results that have been obtained in the auditory domain in which short temporal regularities can be detected unconsciously eliciting a locally generated event-related potential (ERP), termed mismatch negativity (MMN), while detection of long-term regularities depends on conscious perception, which elicits an electrophysiological response known as P300 from a widespread network of brain areas (Bekinschtein et al. 2009; Faugeras et al. 2011).

Together, I propose that these results mesh well with the idea that one of the functions of consciousness is to interpret the world in *long timescales*, bringing together the *now* with the past beyond the simple and automatic input-output relations rooted in unconscious processors, thus allowing for the extraction of more complex and abstract regularities. Brain areas with longer time constants such as the prefrontal cortex (Fuster 1973) would extract the world's statistics from the remote past, creating a model of the world that keeps a stable picture. In contrast, early sensory areas with short time constants act on shorter timescales, sampling the world for alternative interpretations, thus allowing the system to stay tuned to deviations from the long-term statistics (Clifford 2012; Snyder et al. under review).⁴

While previous studies and established experimental paradigms have mostly focused on the "nowness" of conscious perception, it ap-

pears that much remains to be learned about consciousness and its fundamental phenomenological characteristics such as its flow and our sense of stability over time. In fact, considering that much of what we currently know about the NCC stems from "static" paradigms, and by those I mean paradigms that do not take the temporal context in which the stimuli unravel into account and thus only inform us about what has "changed" in consciousness, we in fact only have access to the neural processes related to the *update* of contents in consciousness, while the mechanisms at play in the *maintenance* or continuity of our experience remain obscure (but see Kleinschmidt et al. 2002). The *present* might be known, but the *flow* is still a mystery!

Thus I propose that a full account of consciousness requires a reappraisal of our object of study in which we incorporate the temporal flow of consciousness as another fundamental property that needs to be explained. This calls for a dynamic view in which a train of conscious states (the flow) would be captured as successions of neuronal meta-assemblies, each with a particular relaxation time, followed by phase transitions, which determine the time of emergence, dominance, and dissolution of a state that leads to another perceptual cycle (Melloni & Singer 2010; Varela 1999). In this framework, the rate-limiting factor for the formation of a new meta-assembly would correspond to the time needed to establish stable phase relations; while the different time constants promoting stability vs. change may be implemented by different oscillatory frequency bands, in addition to the intrinsic time window of integration of a given area (Chaudhuri et al. 2014).

In summary, much remains to be discovered about consciousness and its neural correlates, but significant progress has already been made since the seminal paper by Crick & Koch (1990) that got the field going about twenty-five years ago. Victor Lamme's experimental work and theoretical proposals on the role of feedback connections and reentrant activity in conscious perception have been central to bringing us closer to an understanding of the neural processes that allow us to "see". His paper in this volume contains an erudite review

⁴ Similarly, higher areas have larger receptive fields than lower areas, allowing integration over larger regions of space, and are often more broadly tuned (i.e., allow for more variability in the stimulus, e.g., different views of the same object). This resonates well with psychophysical evidence that hysteresis is spatially less specific and more broadly tuned than adaptation (Gepshtein & Kubovy 2005; Knapen et al. 2009).

of the present knowledge against a background of thought provoking hypotheses, e.g., that the function of consciousness is to solve difficult perceptual problems. In Lamme's view, consciousness is there to create, while unconscious processes are there to utilise. In close analogy to any creative process, consciousness in Lamme's framework is slow and takes time and resources to develop. In a way, his proposal is that it is all about distance, or time. This is a powerful intuition, and an idea worth exploring, yet its contribution does not end there—more than that, it serves as a reminder of a central characteristic of consciousness that is not yet fully explored, namely that conscious experience unfolds at a characteristic spatio-temporal scale, and that it is this flow in space/time that brings the strong sense of experiential stability and continuity. The interwoven temporal scales of the flow of consciousness that bring about the “unity of experience” remain the next challenge, and maybe the one that will finally unlock the mystery of consciousness.

3.1 Acknowledgements

This work was supported by a Marie Curie International Outgoing Fellowship of the European Community's Seventh Framework Programme under project number 299372. I am indebted to Caspar M. Schwiedrzik for helpful discussions while writing this commentary, but foremost to Thomas Metzinger and Jennifer Windt for providing a stimulating, open, and alive environment for discussions during the MIND meetings, and also to two anonymous reviewers for their insightful comments.

References

- Almeida, J., Pajtas, P. E., Mahon, B. Z., Nakayama, K. & Caramazza, A. (2013). Affect of the unconscious: Visually suppressed angry faces modulate our decisions. *Cognitive, Affective, & Behavioral Neuroscience*, 13 (1), 94-101. [10.3758/s13415-012-0133-7](https://doi.org/10.3758/s13415-012-0133-7)
- Aru, J., Axmacher, N., Do Lam, A. T., Fell, J., Elger, C. E., Singer, W. & Melloni, L. (2012). Local category-specific gamma band responses in the visual cortex do not reflect conscious perception. *The Journal of Neuroscience*, 32 (43), 14909-14914. [10.1523/JNEUROSCI.2051-12.2012](https://doi.org/10.1523/JNEUROSCI.2051-12.2012)
- Baars, B. J. (2002). The conscious access hypothesis: Origins and recent evidence. *Trends in Cognitive Sciences*, 6 (1), 47-52. [10.1016/S1364-6613\(00\)01819-2](https://doi.org/10.1016/S1364-6613(00)01819-2)
- Bekinschtein, T. A., Dehaene, S., Rohaut, B., Tadel, F., Cohen, L. & Naccache, L. (2009). Neural signature of the conscious processing of auditory regularities. *Proceedings of the National Academy of Sciences of the United States of America*, 106 (5), 1672-1677. [10.1073/pnas.0809667106](https://doi.org/10.1073/pnas.0809667106)
- Burr, D. & Ross, J. (2008). A visual sense of number. *Current Biology*, 18 (6), 425-428. [10.1016/j.cub.2008.02.052](https://doi.org/10.1016/j.cub.2008.02.052)
- Chaudhuri, R., Bernacchia, A. & Wang, X. J. (2014). A diversity of localized timescales in network activity. *Elife*, 3, e01239. [10.7554/eLife.01239](https://doi.org/10.7554/eLife.01239)
- Chopin, A. & Mamassian, P. (2012). Predictive properties of visual adaptation. *Current Biology*, 22 (7), 622-626. [10.1016/j.cub.2012.02.021](https://doi.org/10.1016/j.cub.2012.02.021)
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36 (3), 181-204. [10.1017/S0140525X12000477](https://doi.org/10.1017/S0140525X12000477)
- Clifford, C. W. (2012). Visual perception: knowing what to expect. *Current Biology*, 22 (7), 223-225. [10.1016/j.cub.2012.02.019](https://doi.org/10.1016/j.cub.2012.02.019)
- Cohen, M. X., van Gaal, S., Ridderinkhof, K. R. & Lamme, V. A. (2009). Unconscious errors enhance prefrontal-occipital oscillatory synchrony. *Frontiers in Human Neuroscience*, 3 (54), 1-12. [10.3389/neuro.09.054.2009](https://doi.org/10.3389/neuro.09.054.2009)
- Crick, F. & Koch, C. (1990). Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences*, 2, 263-275.
- de Gardelle, V., Charles, L. & Kouider, S. (2011). Perceptual awareness and categorical representation of faces: Evidence from masked priming. *Consciousness and Cognition*, 20 (4), 1272-1281. [10.1016/j.concog.2011.02.001](https://doi.org/10.1016/j.concog.2011.02.001)

- Dehaene, S. (2014). *Consciousness and the brain: Deciphering how the brain codes our thoughts*. New York, NY: Viking Penguin.
- Dehaene, S. & Changeux, J. P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, 70 (2), 200-227. [10.1016/j.neuron.2011.03.018](https://doi.org/10.1016/j.neuron.2011.03.018)
- Dehaene, S., Naccache, L., Le Clec, H. G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., van de Moortele, P.-F. & Le Bihan, D. (1998). Imaging unconscious semantic priming. *Nature*, 395 (6702), 597-600. [10.1038/26967](https://doi.org/10.1038/26967)
- Del Zotto, M., Deiber, M. P., Legrand, L. B., De Gelder, B. & Pegna, A. J. (2013). Emotional expressions modulate low alpha and beta oscillations in a cortically blind patient. *International Journal of Psychophysiology*, 90 (3), 358-362. [10.1016/j.ijpsycho.2013.10.007](https://doi.org/10.1016/j.ijpsycho.2013.10.007)
- Di Lollo, V., Enns, J. T. & Rensink, R. A. (2000). Competition for consciousness among visual events: The psychophysics of reentrant visual processes. *Journal of Experimental Psychology: General*, 129 (4), 481-507.
- Edelman, G. M. & Tononi, G. (2000). *A universe of consciousness: How matter becomes imagination*. New York, NY: Basic Books.
- Faugeras, F., Rohaut, B., Weiss, N., Bekinschtein, T. A., Galanaud, D., Puybasset, L., Bolgert, F., Sergent, C., Cohen, L., Dehaene, S. & Naccache, L. (2011). Probing consciousness with event-related potentials in the vegetative state. *Neurology*, 77 (3), 264-268. [10.1212/WNL.0b013e3182217ee8](https://doi.org/10.1212/WNL.0b013e3182217ee8)
- Fischer, J. & Whitney, D. (2014). Serial dependence in visual perception. *Nature Neuroscience*, 17 (5), 738-743. [10.1038/nn.3689](https://doi.org/10.1038/nn.3689)
- Friston, K. J. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11 (2), 127-138. [10.1038/nrn2787](https://doi.org/10.1038/nrn2787)
- Friston, K., Adams, R. A., Perrinet, L. & Breakspear, M. (2012). Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology*, 3, 1-20. [10.3389/fpsyg.2012.00151](https://doi.org/10.3389/fpsyg.2012.00151)
- Fuster, J. M. (1973). Unit activity in prefrontal cortex during delayed-response performance: Neuronal correlates of transient memory. *Journal of Neurophysiology*, 36 (1), 61-78.
- Gaillard, R., Dehaene, S., Adam, C., Clémenceau, S., Hasboun, D., Baulac, M., Cohen, L. & Naccache, L. (2009). Converging intracranial markers of conscious access. *PLoS Biology*, 7 (3), e61. [10.1371/journal.pbio.1000061](https://doi.org/10.1371/journal.pbio.1000061)
- Gepshtein, S. & Kubovy, M. (2005). Stability and change in perception: spatial organization in temporal context. *Experimental Brain Research*, 160 (4), 487-495. [10.1007/s00221-004-2038-3](https://doi.org/10.1007/s00221-004-2038-3)
- Greenwald, A. G., Abrams, R. L., Naccache, L. & Dehaene, S. (2003). Long-term semantic memory versus contextual memory in unconscious number processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29 (2), 235-247. [10.1037/0278-7393.29.2.235](https://doi.org/10.1037/0278-7393.29.2.235)
- Hipp, J. F., Engel, A. K. & Siegel, M. (2011). Oscillatory synchronization in large-scale cortical networks predicts perception. *Neuron*, 69 (2), 387-396. [10.1016/j.neuron.2010.12.027](https://doi.org/10.1016/j.neuron.2010.12.027)
- Hock, H. S., Schoner, G. & Hochstein, S. (1996). Perceptual stability and the selective adaptation of perceived and unperceived motion directions. *Vision Research*, 36 (20), 3311-3323. [10.1016/0042-6989\(95\)00277-4](https://doi.org/10.1016/0042-6989(95)00277-4)
- Hohwy, J. (2013). *The predictive mind*. Oxford, UK: Oxford University Press.
- Jardri, R. & Deneve, S. (2013). Circular inferences in schizophrenia. *Brain*, 136 (11), 3227-3241. [10.1093/brain/awt257](https://doi.org/10.1093/brain/awt257)
- Jenkins, R., Beaver, J. D. & Calder, A. J. (2006). I thought you were looking at me: direction-specific aftereffects in gaze perception. *Psychological Science*, 17 (6), 506-513. [10.1111/j.1467-9280.2006.01736.x](https://doi.org/10.1111/j.1467-9280.2006.01736.x)
- Jones, P. D. & Holding, D. H. (1975). Extremely long-term persistence of the McCollough effect. *Journal of Experimental Psychology: Human Perception and Performance*, 1 (4), 323-327. [10.1037/0096-1523.1.4.323](https://doi.org/10.1037/0096-1523.1.4.323)
- King, J. R., Sitt, J. D., Faugeras, F., Rohaut, B., El Karoui, I., Cohen, L. & Dehaene, S. (2013). Information sharing in the brain indexes consciousness in noncommunicative patients. *Current Biology*, 23 (19), 1914-1919. [10.1016/j.cub.2013.07.075](https://doi.org/10.1016/j.cub.2013.07.075)
- Kleinschmidt, A., Büchel, C., Hutton, C., Friston, K. J. & Frackowiak, R. S. (2002). The neural structures expressing perceptual hysteresis in visual letter recognition. *Neuron*, 34 (4), 659-666. [10.1016/S0896-6273\(02\)00694-3](https://doi.org/10.1016/S0896-6273(02)00694-3)
- Knapen, T., Brascamp, J., Adams, W. J. & Graf, E. W. (2009). The spatial scale of perceptual memory in ambiguous figure perception. *Journal of Vision*, 9 (13), 11-12. [10.1167/9.13.16](https://doi.org/10.1167/9.13.16)
- Koch, C. (2004). *The quest for consciousness: A neurobiological approach*. Englewood, CO: Roberts & Company.

- Kouider, S. & Dehaene, S. (2007). Levels of processing during non-conscious perception: A critical review of visual masking. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 362 (1481), 857-875. [10.1098/rstb.2007.2093](#)
- Lamme, V. (2015). The crack of dawn: Perceptual functions and neural mechanisms that mark the transition from unconscious processing to conscious vision. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Lin, Z. & He, S. (2009). Seeing the invisible: The scope and limits of unconscious processing in binocular rivalry. *Progress in Neurobiology*, 87 (4), 195-211. [10.1016/j.pneurobio.2008.09.002](#)
- Mayer, A., Schwiedrzik, C. M., Singer, W. & Melloni, L. (in preparation). *Expectations sparsify networks for letter recognition*.
- Melloni, L., Molina, C., Pena, M., Torres, D., Singer, W. & Rodriguez, E. (2007). Synchronization of neural activity across cortical areas correlates with conscious perception. *The Journal of Neuroscience*, 27 (11), 2858-2865. [10.1523/JNEUROSCI.4623-06.2007](#)
- Melloni, L., Schwiedrzik, C. M., Rodriguez, E. & Singer, W. (2009). (Micro)Saccades, corollary activity and cortical oscillations. *Trends in Cognitive Sciences*, 13 (6), 239-245. [10.1016/j.tics.2009.03.007](#)
- Melloni, L., Schwiedrzik, C. M., Muller, N., Rodriguez, E. & Singer, W. (2011). Expectations change the signatures and timing of electrophysiological correlates of perceptual awareness. *The Journal of Neuroscience*, 31 (4), 1386-1396. [10.1523/JNEUROSCI.4570-10.2011](#)
- Melloni, L. & Rodriguez, E. (2007). Non-perceived stimuli elicit local but not large-scale neural synchrony. *Perception*, 36 (ECP Abstract Supplement)
- Melloni, L. & Singer, W. (2010). Distinct characteristics of conscious experience are met by large-scale neuronal synchronization. In E. K. Perry, D. Collerton, F. E. N. LeBeau & H. Ashton (Eds.) *New horizons in the neuroscience of consciousness* (pp. 17-28). Amsterdam, NL: John Benjamins.
- Moca, V. V., Tincas, I., Melloni, L. & Muresan, R. C. (2011). Visual exploration and object recognition by lattice deformation. *PLoS One*, 6 (7), e22831. [10.1371/journal.pone.0022831](#)
- Mudrik, L., Faivre, N. & Koch, C. (2014). Information integration without awareness. *Trends in Cognitive Sciences*, 18 (8), 414-421. [10.1016/j.tics.2014.04](#)
- Murray, S. O., Schrater, P. & Kersten, D. (2004). Perceptual grouping and the interactions between visual cortical areas. *Neural Networks*, 17 (5-6), 695-705. [10.1016/j.neunet.2004.03.010](#)
- Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83 (4), 435-450. [10.2307/2183914](#)
- Pellicano, E. & Burr, D. (2012). When the world becomes 'too real': A Bayesian explanation of autistic perception. *Trends in Cognitive Sciences*, 16 (10), 504-510. [10.1016/j.tics.2012.08.009](#)
- Purkinje, J. E. (1820). Beiträge zur näheren Kenntnis des Schwindels aus heautognostischen Daten. *Medizinische Jahrbücher des kaiserl.-königl. österreichischen Staates*, 6, 79-125.
- Schroeder, C. E., Wilson, D. A., Radman, T., Scharfman, H. & Lakatos, P. (2010). Dynamics of active sensing and perceptual selection. *Current Opinion in Neurobiology*, 20 (2), 172-176. [10.1016/j.conb.2010.02.010](#)
- Schwartz, O., Hsu, A. & Dayan, P. (2007). Space and time in visual context. *Nature Reviews Neuroscience*, 8 (7), 522-535. [10.1038/nrn2155](#)
- Schwiedrzik, C. M., Singer, W. & Melloni, L. (2009). Sensitivity and perceptual awareness increase with practice in metacontrast masking. *Journal of Vision*, 9 (10), 11-18. [10.1167/9.10.18](#)
- (2011). Subjective and objective learning effects dissociate in space and in time. *Proceedings of the National Academy of Sciences of the United States of America*, 108 (11), 4506-4511. [10.1073/pnas.1009147108](#)
- (2011). Subjective and objective learning effects dissociate in space and in time. *Proceedings of the National Academy of Sciences of the United States of America*, 108 (11), 4506-4511. [10.1073/pnas.1009147108](#)
- Schwiedrzik, C. M., Ruff, C. C., Lazar, A., Leitner, F. C., Singer, W. & Melloni, L. (2014). Untangling perceptual memory: Hysteresis and adaptation map into separate cortical networks. *Cerebral Cortex*, 24 (5), 1152-1164. [10.1093/cercor/bhs396](#)
- Seth, A. K., Suzuki, K. & Critchley, H. D. (2011). An interoceptive predictive coding model of conscious presence. *Front in Psychology*, 2 (395), 1-16. [10.3389/fpsyg.2011.00395](#)
- Snyder, J., Schwiedrzik, C. M., Vitela, D. & Melloni, L. (forthcoming). *How previous experience shapes perception across sensory modalities*.
- Thompson, E. & Varela, F. J. (2001). Radical embodiment: Neural dynamics and consciousness. *Trends in Cognitive Sciences*, 5 (10), 418-425. [10.1016/S1364-6613\(00\)01750-2](#)
- Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, 5 (45), 1-22. [10.1186/1471-2202-5-42](#)

- Tononi, G. & Koch, C. (2008). The neural correlates of consciousness: An update. *Annals of the New York Academy of Sciences*, 1124 (1), 239-261. [10.1196/annals.1440.004](https://doi.org/10.1196/annals.1440.004)
- Treisman, M. (1984). A theory of criterion setting: An alternative to the attention band and response ratio hypotheses in magnitude estimation and cross-modality matching. *Journal of Experimental Psychology: General*, 113 (3), 443-463. [10.1037/0096-3445.113.3.443](https://doi.org/10.1037/0096-3445.113.3.443)
- van Leeuwen, T. M., Wibral, M., Sauer, A., Uhlhaas, P., Singer, W. & Melloni, L. (2013). Neural synchronization during bottom-up and top-down visual processing in grapheme- color synesthetes and schizophrenia patients. *Poster at the 43rd Meeting of the Society for Neuroscience (SfN), San Diego, USA*.
- Varela, F. (1999). The specious present: A neurophenomenology of time consciousness. In J. Petitot, J. Varela, J.-M. Roy & B. Pachoud (Eds.) *Naturalizing phenomenology* (pp. 266-314). Stanford, CA: Stanford University Press.
- Varela, F., Lachaux, J. P., Rodriguez, E. & Martinerie, J. (2001). The brainweb: Phase synchronization and large-scale integration. *Nature Reviews Neuroscience*, 2 (4), 229-239. [10.1038/35067550](https://doi.org/10.1038/35067550)
- von Foerster, H. (1984). On constructing a reality. In P. Watzlawick (Ed.) *The invented reality: How do we know what we believe we know* (pp. 41-62). New York: W.W.Norton & Co.
- von Helmholtz, H. (1962). *Handbuch der physiologischen Optik*. New York, NY: Dover.
- Webster, M. A., Kaping, D., Mizokami, Y. & Duhamel, P. (2004). Adaptation to natural facial categories. *Nature*, 428 (6982), 557-561. [10.1038/nature02420](https://doi.org/10.1038/nature02420)

Predictive Coding Is Unconscious, so that Consciousness Happens *Now*

A Reply to Lucia Melloni

Victor Lamme

Conscious percepts depend strongly on past events. Expectations, primes, and prior experiences all shape the percept we have at any moment in time. Yet does this imply that conscious experience should be viewed as extended in time—as “flowing”—instead of as just happening now?

Keywords

Bayesian framework | Inference | Predictive coding | Snapshot vision | Spatial integration | Stream of consciousness | Temporal integration

Author

[Victor Lamme](#)

Victorlamme@gmail.com

Universiteit van Amsterdam

Amsterdam, Netherlands

Commentator

[Lucia Melloni](#)

lucia.melloni@brain.mpg.de

Max Planck Institute for Brain Research

Frankfurt a. M., Germany

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 To infer or to integrate, that is the question

In her commentary, Lucia Melloni argues that consciousness unfolds in time: there is a stream of consciousness. What I see now is intricately linked to what I have seen before. And what I see now is what I expect to see—much along the lines of predictive coding. A full understanding of consciousness should not neglect this point. There is even a stronger claim that somehow the process of inference over time is crucial to understanding consciousness.

I appreciate the boldness of linking the framework of Bayesian predictive coding to specific stages in the process of generating consciousness:

One promising framework within which the influence of previous experience can be understood is the Bayesian framework. When applied to perception, each mathematically-formulated ingredient of this framework can be assigned a percep-

tual counterpart, with previous experience referring to the prior, the current moment referring to the likelihood, unconscious inference referring to Bayes rule (which combines the prior with the likelihood in an optimal way), and the result—our perception—referring to (the peak of) the posterior distribution. ([Melloni this collection](#), p. 4)

To my knowledge, this is the first time this has been so explicitly laid out—writers on predictive coding thus far have always stayed a little vague on where exactly consciousness sits in the Bayesian framework.

Yet at the same time, there is the suggestion of long temporal range integration being the key ingredient of consciousness:

Event-objects of the conscious mind thus per definition unfold in time and we also act in time: we make use of current and previous input to figure out the most appropriate response predicting their consequences. There is thus a continuum of interdependencies along the time dimension whereby every past moment is integrated with the present and projected into the future, giving rise to the flow of consciousness. The same way we have been thinking about the integration of multiple sources of information within a given moment of time, such as multiple features of a single object, there is thus integration across time. ([Melloni this collection](#), pp. 7-8)

This makes intuitive sense, particularly in the case of moving objects, such as the tennis ball Melloni uses as an example. Indeed it is hard—if not impossible—to pinpoint the exact *now* of conscious experience of such a ball.¹

Yet the two points seem contradictory. In the Bayesian predictive coding framework, consciousness is the *result* of the unconscious inferential processes. Previous knowledge and

experience (the priors) play an important role, but they are combined with current input to produce the posterior, which is conscious sensation. In the second account, however, consciousness seems to be something that is stretched out over time, so that both prior and posterior are smelted into a “flow” of consciousness. I find it hard to reconcile these two views.

2 The latency of visual consciousness is variable

Melloni discusses some impressive experiments that show the crucial importance of prior information and expectation in shaping or simply altering conscious experience (and her example, figure 1, is enlightening and flattering at the same time). In all these cases, however, consciousness is portrayed as the *outcome* or *result* of an otherwise unconscious inferential process. The result may come earlier or later, as in the experiment on letter priming that Melloni describes, resulting in earlier (200ms) or later (300ms) electrophysiological correlates of conscious recognition depending on the presence or absence of appropriate priors. Further experiments are discussed, showing that neural correlates of consciousness may shift (neural) location, depending on expectation and priors. Yet still, the end result—consciousness—occurs at the end of a cascade of neural operations. Consciousness, in this account, may occur at variable moments and locations, but *moments* they are.

These results complement earlier findings that the latency of recurrent processing—and hence the emergence of a conscious sensation—may vary. [Super et al. \(2001\)](#) showed that degrading stimulus quality may increase the latency of recurrent signals to V1 in the monkey visual cortex (see figure 5c of [Supèr et al. 2001](#)), and that this affects the latency of behavioral responses of animals that are consciously reporting the presence or absence of the stimuli. Latency of recurrent signals may also vary spontaneously between trials, which correlates with the latency of memory-guided—but not reflexive—saccades to the targets

¹ Although some have argued that consciousness unfolds in time as a succession of static frames, more or less like the single frames of a movie—even at specific frequencies, namely 10Hz and 40Hz ([Van Rullen & Koch 2003](#)).

that elicit these recurrent signals (Supèr et al. 2004). In humans, the latencies of electrophysiological correlates of recurrent processing also vary, either spontaneously or depending on stimulus properties (Jolij et al. 2011), or depending on the IQ of the subject (Jolij et al. 2007). Likewise, this has consequences for the latency of conscious sensations. The Jolij 2011 study, for example, found that variations in the latency of recurrent EEG signals covary with variations in subjective simultaneity of the stimuli evoking these signals.² These results invariably imply that consciousness arises at a particular *moment* in time. That moment may vary from stimulus to stimulus, from trial to trial, from person to person, from prior to prior. But nothing is flowing or stretched out over time.

3 Consciousness is not streaming, but taking snapshots

One may argue that these findings are all obtained with stimuli that are presented *de novo*, using the classic stimulus-onset paradigms. In normal vision, things don't suddenly appear out of nowhere. Or do they? We naturally make about three saccadic eye movements per second, and each time the eye lands on a "new" scene which is—from a retinotopic point of view—radically different from the previous one. In between, we are blind due to saccadic suppression. Moreover, little information seems to be transferred from one view to the next, although some (attended) neural representations seem to be *remapped* across saccades (see Bays & Husain 2007, for an overview of trans-saccadic memory and neural remapping). Such a remapping may allow for a more efficient saccade from one object to the next, when both were already present before the first saccade was made. The predictive coding framework seems to re-emerge in this context: objects that were present or attended on a first fixa-

tion form a sort of prior for the representation that is built during the second fixation (which may then arise more rapidly).

Melloni further claims that previous experience has different effects on what is perceived now depending on the temporal interval between prior and current experience. Bistable percepts show hysteresis or adaptation depending on these temporal intervals, or depending on whether the previous experience was conscious or not. But again, I fail to see how these findings support the idea that consciousness is stretched out over time instead of just happening *now*.³

So I appreciate the importance of the predictive coding framework. Previous experience plays a very important role in the conscious sensations we have, and the why and how of this is extremely important for fully understanding vision. But these contributions are unconscious. Consciousness happens now, and its neural correlates are likewise limited in time. Consciousness of the past we call memory.

Acknowledgements

This work was supported by an advanced investigator grant from the ERC.

² For this reason, I don't quite understand why Melloni suggests that I am claiming that consciousness arises at a particular and *fixed* moment in time. My claim is only that it comes *after* feedforward processing, and as soon as recurrent processing emerges—which may vary.

³ Of course there are visual percepts that are more or less defined by their temporal sequence, the prime example being motion. But this does not imply that the perception of motion is flowing. The first thing the brain does in detecting motion is to convert the flow of motion into a discrete and momentary signal, indistinguishable from how the brain represents other features such as orientation, color, or shape. As a result we see motion now, and instantaneously, which is also crucial for our survival: perceiving something moving in the shadows of a bush (e.g., a snake) needs to be translated into action as soon as possible (e.g., running away). No time for any flow there.

References

- Bays, P. M. & Husain, M. (2007). Spatial remapping of the visual world across saccades. *Neuroreport*, 18 (12), 1207-1213. [10.1097/WNR.0b013e328244e6c3](https://doi.org/10.1097/WNR.0b013e328244e6c3)
- Jolij, J., Huisman, D., Scholte, H. S., Hamel, R., Kemner, C. & Lamme, V. A. F. (2007). Processing speed in recurrent visual networks correlates with general intelligence. *Neuroreport*, 18 (1), 39-43. [10.1097/01.wnr.0000236863.46952.a6](https://doi.org/10.1097/01.wnr.0000236863.46952.a6)
- Jolij, J., Scholte, H. S., Van Gaal, S., Hodgson, T. L. & Lamme, V. A. F. (2011). Act quickly, decide later: Long-latency visual processing underlies perceptual decisions but not reflexive behavior. *Journal of Cognitive Neuroscience*, 23 (12), 3734-3745. [10.1162/jocn_a_00034](https://doi.org/10.1162/jocn_a_00034)
- Melloni, L. (2015). Consciousness as inference in time. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Supèr, H., Spekreijse, H. & Lamme, V. A. F. (2001). Two distinct modes of sensory processing observed in monkey primary visual cortex (V1). *Nature Neuroscience*, 4, 304-310. [10.1038/85170](https://doi.org/10.1038/85170)
- Supèr, H., Van der Togt, C., Spekreijse, H. & Lamme, V. A. F. (2004). Correspondence of presaccadic activity in the monkey primary visual cortex with saccadic eye movements. *Proceedings of the National Academy of Sciences of the United States of America*, 101 (9), 3230-3235. [10.1073/pnas.0400433101](https://doi.org/10.1073/pnas.0400433101)
- Van Rullen, R. & Koch, C. (2003). Is perception discrete or continuous? *Trends in Cognitive Sciences*, 7 (5), 207-213. [10.1016/S1364-6613\(03\)00095-0](https://doi.org/10.1016/S1364-6613(03)00095-0)

Vestibular Contributions to the Sense of Body, Self, and Others

Bigna Lenggenhager & Christophe Lopez

There is increasing evidence that vestibular signals and the vestibular cortex are not only involved in oculomotor and postural control, but also contribute to higher-level cognition. Yet, despite the effort that has recently been made in the field, the exact location of the human vestibular cortex and its implications in various perceptual, emotional, and cognitive processes remain debated. Here, we argue for a vestibular contribution to what is thought to fundamentally underlie human consciousness, i.e., the bodily self. We will present empirical evidence from various research fields to support our hypothesis of a vestibular contribution to aspects of the bodily self, such as basic multisensory integration, body schema, body ownership, agency, and self-location. We will argue that the vestibular system is especially important for global aspects of the self, most crucially for implicit and explicit spatiotemporal self-location. Furthermore, we propose a novel model on how vestibular signals could not only underlie the perception of the self but also the perception of others, thereby playing an important role in embodied social cognition.

Keywords

Agency | Bodily self | Consciousness | Interoception | Multisensory integration | Ownership | Self-location | Vestibular system

Authors

[Bigna Lenggenhager](#)

bigna.lenggenhager@gmail.com

University Hospital

Zurich, Switzerland

[Christophe Lopez](#)

christophe.lopez@univ-amu.fr

CNRS and Aix Marseille Université

Marseille, France

Commentator

[Adrian Alsmith](#)

adrianjtalsmith@gmail.com

Københavns Universitet

Copenhagen, Denmark

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität

Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University

Melbourne, Australia

1 Introduction

There is an increasing interest from both theoretical and empirical perspectives in how the central nervous system dynamically represents the body and how integrating bodily signals arguably gives rise to a stable sense of self and self-consciousness (e.g., [Blanke & Metzinger 2009](#); [Blanke 2012](#); [Gallagher 2005](#); [Legrand 2007](#); [Metzinger 2007](#); [Seth 2013](#)). Discussion of the “bodily self”—which is thought to be largely pre-reflective and thus independent of

higher-level aspects such as language and cognition—has played an important role in various theoretical views (e.g., [Alsmith 2012](#); [Blanke 2012](#); [Legrand 2007](#); [Metzinger 2003](#); [Metzinger 2013](#); [Serino et al. 2013](#)). For example in the conceptualisation of minimal phenomenal selfhood (MPS), which constitutes the simplest form of self-consciousness, [Blanke & Metzinger \(2009\)](#) suggested three key features of the MPS: a globalized form of identification with the body

as a whole (as opposed to ownership for body parts), self-location—by which one’s self seems to occupy a certain volume in space at a given time—and a first-person perspective that normally originates from this volume of space.¹ In recent years, an increasing number of studies has tried to manipulate and investigate these aspects of the minimal self as well as other aspects of the bodily self empirically. This chapter aims to show that including the oft-neglected vestibular sense of balance (Macpherson 2011) into this research might enable us to enrich and refine such empirical research as well as its theoretical models and thus gain further insights into the nature of the bodily self. We agree with Blanke & Metzinger (2009) that self-identification, self-location, and perspective are fundamental for the sense of a bodily self and argue that exactly these components are most strongly influenced by the vestibular system. Yet, we additionally want to stress that the phenomenological sense of a bodily self is—at least in a normal conscious waking state—much richer and involves various fine-graded and often fluctuating bodily sensations. We will thus also describe how the vestibular system might contribute to these (maybe not minimal) aspects of bodily self (e.g., the feeling of agency).

The aim of this book chapter is thus to combine findings from human and non-human animal vestibular research with the newest insights from neuroscientific investigations of the sensorimotor foundations of the sense of self. We present several new experimentally testable hypotheses out of this convergence, especially regarding the relation between vestibular coding and the sense of self-location. We first describe the newest advances in the field of experimental studies of the bodily self (section 2) and give a short overview of vestibular processing and multisensory integration along the vestibulo-

thalamo-cortical pathways (section 3). In section 4, we present several lines of evidence and hypotheses on how the vestibular system contributes to various bodily experiences thought to underpin our sense of bodily self. We conclude this section by suggesting that the vestibular system not only contributes to the sense of self, but may also play a significant role in self-other interactions and social cognition.

2 Multisensory mechanisms underlying the sense of the body and self

How the body shapes human conscious experience is an old and controversial philosophical debate. Yet, recent theories converge on the importance of sensory and motor bodily signals for the experience of a coherent sense of self and hence for self-consciousness in general (Berlucchi & Aglioti 2010; Bermúdez 1998; Blanke & Metzinger 2009; Carruthers 2008; Gallagher 2000; Legrand 2007; Metzinger 2007; Tsakiris 2010). Even the emergence of self-consciousness in infants has been linked to their ability to progressively detect intermodal congruence (e.g., Bahrick & Watson 1985; Filippetti et al. 2013; Rochat 1998).² The assumption that multisensory integration of bodily signals underpins the sense of a bodily self has opened up—next to clinical research—a broad and exciting avenue of experimental investigations in psychology and cognitive neuroscience as well as interdisciplinary projects integrating philosophy and neuroscience. Experiments in these fields typically provided participants with conflicting information about certain aspects of their body and assessed how it affected implicit and explicit aspects of the body and self. The first anecdotal evidence of an altered sense of self through exposure to a multisensory conflict dates back at least to the nineteenth century with the work of Stratton (1899). More systematic, well-controlled paradigms from experimental psychology have gained tremendous influence since the first description of the *rubber*

1 Jennifer Windt (2010) suggested, based on dream research, an even more basic form of minimal phenomenal selfhood, which she defined as a “sense of immersion or of (unstable) location in a spatiotemporal frame of reference”, thus not needing a global full-body representation (see also Metzinger 2013, 2014 for an interesting discussion of this view). We believe that for this more basic sense of a self especially, the vestibular system should be of importance, as a vestibular signal unambiguously tells us that our self was moving (i.e., change in self-location and perspective) without an actual sensation from the body (i.e., a specific body location as it is the case in touch, proprioception, or pain).

2 It is interesting to note for the frame of this chapter that these authors describe the importance of the detection of coherence of all self-motion specific information (including the vestibular system), despite the fact that their experimental setup involved only proprioceptive and visual information (leg movements in a sitting position).

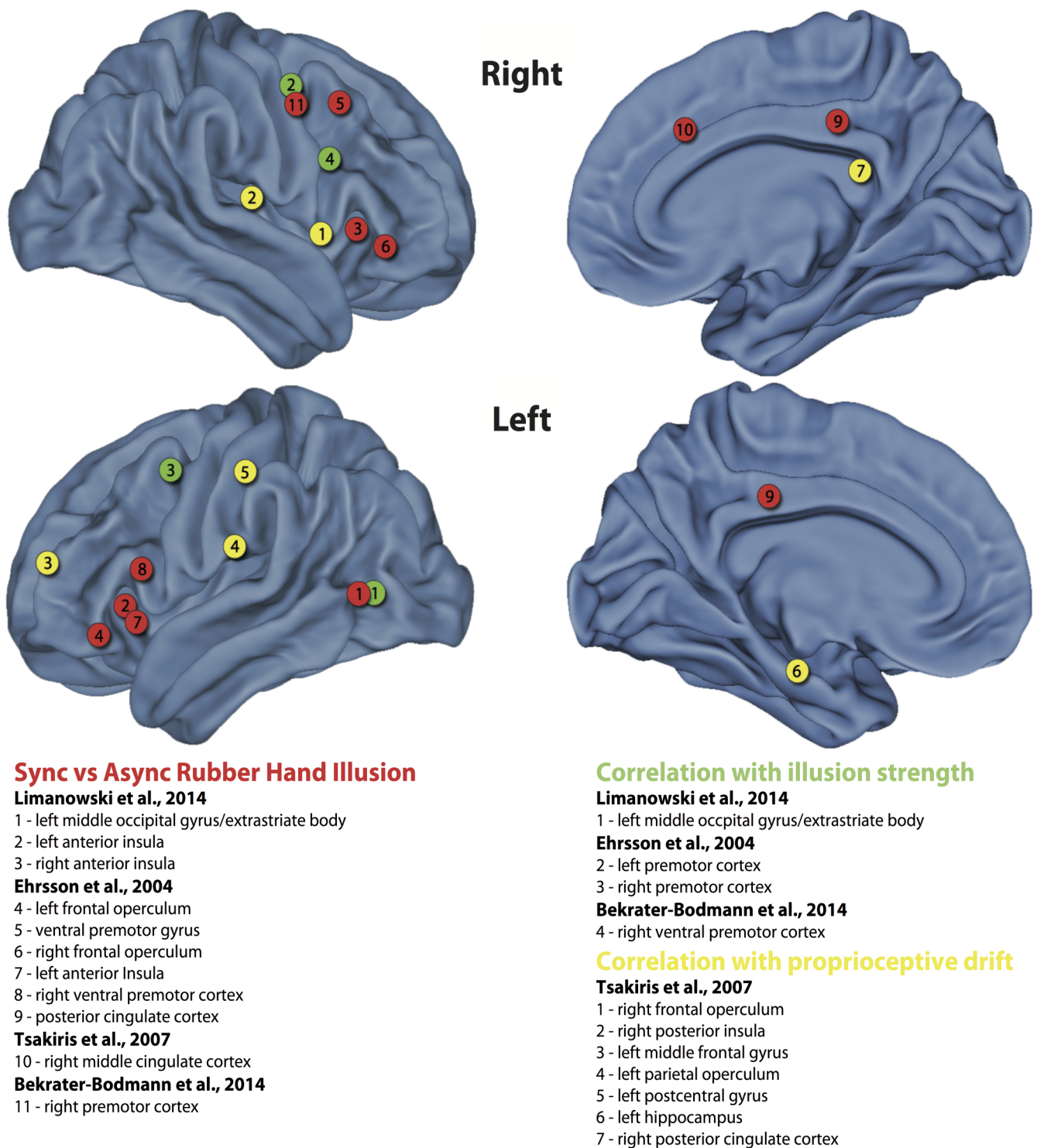


Figure 1: An overview of brain imaging studies of the rubber hand illusion (Bekrater-Bodmann et al. 2014; Ehrsson et al. 2004; Limanowski et al. 2014; Tsakiris et al. 2006). Red circles indicate significant brain activation in the comparison of synchronous visuo-tactile stimulation (illusion condition) to the control asynchronous visuo-tactile stimulation. Green circles indicate brain areas where the hemodynamic response correlates with the strength of the rubber hand illusion. Yellow circles indicate areas that significantly correlate with the proprioceptive drift. For the generation of the figure, MNI coordinates were extracted from the original studies and mapped onto a template with caret (<http://www.nitrc.org/projects/caret/> (van Essen et al. 2001)).

hand illusion seventeen years ago (Botvinick & Cohen 1998). Since then, different important components underlying the bodily self have been identified, described, and experimentally modified. Most prominently: *self-location*—the feeling of being situated at a single location in space; *first-person perspective*—the centeredness of the subjective multidimensional and multimodal experiential space upon one's own body (Vogeley & Fink 2003); *body ownership*—the sense of ownership of the body (Blanke & Metzinger 2009; Serino et al. 2013); and *agency*—the sense of being the agent of one's own actions (Jeannerod 2006). In this section, we briefly describe these components of the bodily self as well as experimental paradigms that allow their systematic manipulation and investigation of their underlying neural mechanisms. Later, in section 4, we will describe how and to what extent vestibular signals might influence these components as well as their underlying multisensory integration.

2.1 Ownership, self-location, and the first-person perspective

2.1.1 Body part illusions

Both ownership and self-location³ have traditionally been investigated in healthy participants using the rubber hand illusion paradigm (Botvinick & Cohen 1998). Synchronous stroking of a hidden real hand and a seen fake hand in front of a participant causes the fake hand to be self-attributed (i.e., quantifiable subjective change in ownership) and the real hand to be mis-localized towards the rubber hand (i.e., objectively quantifiable change in self-location). During the last ten years, various other correlates of the illusion have been described. For example, illusory ownership for a rubber hand is accompanied by a reduction of the skin temperature of the real hand (Moseley et al. 2008), an increased skin conductance and activity in pain-related neural networks in response to a threat toward the rubber hand (Armel &

Ramachandran 2003; Ehrsson et al. 2007), and increased immune response to histamine applied on the skin of the real hand (Barnsley et al. 2011). Several variants of the illusion have been established using conflicts between tactile and *proprioceptive information*,⁴ between visual and nociceptive information (Capelari et al. 2009), between visual and *interoceptive information*, and between visual and motor information (Tsakiris et al. 2007). All these multisensory manipulations have in common that they can induce predictable changes in the implicit and explicit sense of a bodily self. Yet, the question of what components of the bodily self are really altered during such illusions and how the various measures relate to them is still under debate. Longo et al. (2008) used a psychometric analysis of an extended questionnaire presented after the induction of the rubber hand illusion to identify three components of the illusion: (1) *ownership*, i.e., the perception of the rubber hand as part of oneself; (2) *location*, i.e., the localization of one's own hand or of touch applied to one's own hand in the position of the rubber hand; and (3) *sense of agency*, i.e., the experience of control over the rubber hand. These different components seem also to be reflected in differential neural activity as revealed by recent functional neuroimaging studies.⁵

Figure 2 summarizes the main brain regions found to be involved in the rubber hand illusion during functional magnetic resonance imaging (fMRI) or positron emission tomography (PET) studies (Bekrater-Bodmann et al. 2014; Ehrsson et al. 2004; Limanowski et al. 2014; Tsakiris et al. 2006). The activation patterns depend on how the illusion was quantified. The pure contrast of the illusion condition (i.e., synchronous stroking) to the control condition reveals a network including the insular, cingulate, premotor, and lateral occipital (extrastriate body area) cortex. Areas in which haemodynamic responses correlate with the strength of illusory ownership include the premotor cortex

³ This component is in such context usually termed self-location, but a more accurate formulation is “body part location with respect to the self” (Blanke & Metzinger 2009; Lenggenhager et al. 2007).

⁴ Proprioception classically refers to information about the position of body segments originating from muscle spindles, articular receptors, and Golgi tendon organs, while interoception refers to information originating from internal organs such as the heart, gastrointestinal tract, and bladder.

⁵ The sense of agency has not yet been investigated using neuroimaging studies in the context of the rubber hand illusion.

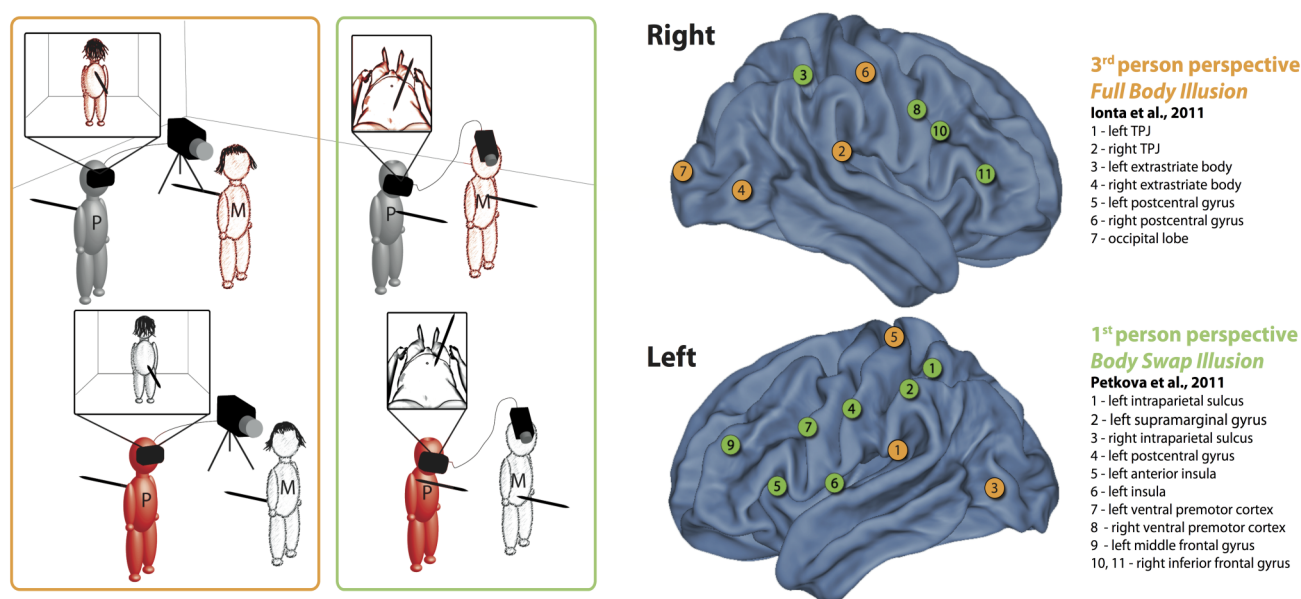


Figure 2: A comparison of brain activity associated with two illusions targeting the manipulation of more global aspects of the bodily self, i.e., the full body illusion (Lenggenhager et al. 2007, setup in orange frame) and the body swap illusion (Petkova & Ehrsson 2008, setup in green frame). In both variants of the illusion, synchronous stroking of one's own body and the seen mannequin led to self-identification with the latter (locus of self-identification is indicated in red colour). Two recent fMRI studies using either the full body illusion (Ionta et al. 2011 in orange circles) or the body swap illusion (Petkova et al. 2011, in green circles) are compared and plotted. Only areas significantly more activated during synchronous visuo-tactile stimulation (illusion condition), as compared to control conditions, are shown. For the generation of the figure, MNI coordinates were extracted from the original studies and mapped onto a template with caret (<http://www.nitrc.org/projects/caret/>). Adapted from Serino et al. 2013, Figure 2.

and extrastriate body area, whereas illusory mis-localization of the physical hand (referred to as “proprioceptive drift”) correlates particularly with responses in the right posterior insula, right frontal operculum, and left middle frontal gyrus (see figure 1 for the detailed list). The fact that different brain regions are involved in illusory ownership and mis-localization of the physical hand provides further evidence for distinct sub-components underlying the bodily self.

2.1.2 Full-body illusions

Several authors claimed that research on body part illusions is unable to provide insight into the mechanisms of global aspects of the bodily self, such as self-identification with a body as a whole, self-location in space, and first-person perspective (e.g., Blanke & Metzinger 2009; Blanke 2012; Lenggenhager et al. 2007). Thus, empirical studies have more recently adapted

the rubber hand illusion paradigm to a *full-body illusion* paradigm where the whole body (instead of just a body part) is seen using video-based techniques and virtual reality.

Two main versions of multisensory illusions targeting more global aspects of the self have been used (but see also Ehrsson 2007), one in which the participants saw the back-view of their own body (or a fake body) in front of them as if it were seen from a third-person perspective (full-body illusion [see figure 2, orange frame]; Lenggenhager et al. 2007) and one in which a fake body was seen from a first-person perspective (body swap illusion [see figure 2 green frame; Petkova & Ehrsson 2008]). In both versions of the illusion, synchronous visuo-tactile stroking of the fake and the real body increased self-identification (i.e., full-body ownership)⁶ with a virtual or fake body as compared

⁶ While these experiments are targeting illusory full-body ownership, it has recently been criticized (Smith 2010; see also Metzinger 2013) that it has not empirically been shown that it really

to asynchronous stroking. Importantly, it has been argued that only the former is associated with a change in self-location⁷ (Aspell et al. 2009; Lenggenhager et al. 2007; Lenggenhager et al. 2009) and in some cases with a change in the direction of the first-person visuo-spatial perspective (Ionta et al. 2011; Pfeiffer et al. 2013).

A recent psychometric approach identified three components of the bodily self in a full-body illusion set up: bodily self-identification, space-related self-perception, which is closely linked to the feeling of presence in a virtual environment (see section 4.5.1.3), and agency (Dobricki & de la Rosa 2013). Again, these sub-components seem to rely on different brain mechanisms. Figure 2 contrasts two recent brain imaging studies using full-body illusions (see Serino et al. 2013, for a more thorough comparison). While self-identification with a fake body seen from a first-person perspective is associated with activity in premotor areas (Petkova et al. 2011), changes in self-location and visuo-spatial perspective are associated with activity in the temporo-parietal junction (TPJ) (Ionta et al. 2011). The TPJ is a region located close to the parieto-insular vestibular cortex (see section 3.2.3), suggesting that the vestibular cortex might play a role in the experienced self-location and visuo-spatial perspective, as we will elaborate on in the following sections.

2.2 Agency

Agency, the feeling that one is initiating, executing, and controlling one's own volitional actions, has been described as another key aspect of the bodily self and self-other discrimination (Gallagher 2000; Jeannerod 2006; Tsakiris et al. 2007). Experimental investigations of the sense of agency started in the 1960s with a study by Nielsen (1963). In this seminal study,

affects the full body (as opposed to just certain body parts). We agree that this argument is justified and that further experiments are needed to address this issue (see also Lenggenhager et al. 2009).

⁷ Similarly to the rubber hand illusion, changes in self-location and self-identification have been associated with physiological changes such as increased pain thresholds, decreased electrodermal response to pain (Romano et al. 2014), and decreased body temperature (Salmom et al. 2013).

as well as in follow-up studies, a spatial or a temporal bias was introduced between a physical action (e.g., reaching movement toward a target) and the visual feedback from this action (Farrer et al. 2003b; Fournier & Jeannerod 1998). These studies measured the degree of discrepancy for which the movement is still self-attributed. Theories of the sense of agency have mostly been based on a “forward model,” which has been defined in a predictive coding framework (Friston 2012). The forward model uses the principle of the *efference motor copy*, which is a copy from the motor commands predicting the sensory consequences of an action. Such efference copies allow the brain to distinguish self-generated actions from externally generated actions (Wolpert & Miall 1996). This idea is supported by a large body of empirical evidence showing that the sense of agency increases with increasing congruence of predicted and actual sensory input (e.g., Farrer et al. 2003a; Fournier et al. 2001). Neurophysiological and brain imaging studies showed a reduction of activation in sensory areas in response to self-generated, as compared to externally generated, movements (e.g., Gentsch & Schütz-Bosbach 2011). As well as suppression of activity in specific sensory areas, agency has also been linked to activity in a large network including the ventral premotor cortex, supplementary motor area, cerebellum, dorsolateral prefrontal cortex, posterior parietal cortex, posterior superior temporal sulcus, angular gyrus, and the insula (David et al. 2006; Farrer et al. 2008; Farrer et al. 2003a).

While studies on agency have almost exclusively investigated agency for arm and hand movements, a recent study has addressed “full-body agency” during locomotion using full-body tracking and virtual reality (Kannape et al. 2010). As the vestibular system is importantly involved in locomotion, we will argue for a strong implication of the vestibular system in full-body agency during locomotion (see section 4.4).

3 The vestibular system

In this section, we describe the basic mechanisms of the peripheral and central vestibular

system for coding self-motion and self-orientation, as we believe that these aspects are crucial bases for a sense of the bodily self. It is, however, beyond the scope of this paper to provide a comprehensive description of the vestibular system anatomy and physiology, and the reader is referred to recent review articles (e.g., Angelaki & Cullen 2008; Lopez & Blanke 2011).

3.1 Peripheral mechanisms

The peripheral vestibular organs in the inner ear contain sensors detecting three-dimensional linear motions (two otolith organs) and angular motions (three semicircular canals). The characteristic of these sensors is that they are *inertial sensors*, a type of accelerometers and gyroscopes found in inertial navigation systems. When an individual turns actively his or her head, or when the head is moved passively (e.g., in a train moving forward), the head acceleration is transmitted to the vestibular organs. Head movements create inertial forces—due to the inertia of the otoconia, the small crystals of calcium carbonate above the otolith organs, and to the inertia of the endolymphatic fluid in the semicircular canals—inducing an activation or inactivation of the vestibular sensory hair cells.

It is important to note here that the neural responses of the vestibular sensory hair cells depend on the direction of head movements with respect to head-centred inertial sensors and not with respect to any external reference. For this reason, the vestibular system enables the coding of *absolute* head motion in a *head-centred reference frame* (Berthoz 2000). This way of coding body motion differs from the motion coding done by other sensory systems. The coding by the visual, somatosensory, and auditory system is ambiguous because these sensory systems detect a body motion *relative to* an external reference, or the motion of an external object with respect to the body. For example, the movement of an image on the retina can be interpreted either as a motion of the body with respect to the visual surrounding, or as a motion of the visual scene in front of a static observer (e.g., Dichgans & Brandt 1978), leading to an am-

biguous sense of ownership for the movement. Similarly, if a subject detects changes of pressures applied to his skin (e.g., under his foot soles), this can be related either to a body movement, with respect to the surface on which he is standing, or to the movement of this surface on his skin (Kavounoudias et al. 1998; Lackner & DiZio 2005). Similar observations have been made in the auditory system and illusory sensations of body motion have been evoked by rotating sounds (Väljamäe 2009). By contrast, a vestibular signal is a non-ambiguous neural signal that the head moved or has been moved; thus there is no ambiguity regarding whether the own body moved or the environment moved. It should, however, be noted that the vestibular information on its own does not distinguish between passive or active movements of the subject's whole body (i.e., the self-motion associated with the feeling of agency; see also section 4.4).⁸

The otolith organs are not only activated by head translations, such as those produced by a train moving forward or by an elevator moving upward, but also by Earth's gravitational pull. Otolith receptors are sensitive to *gravito-inertial forces* (Angelaki et al. 2004; Fernández & Goldberg 1976) and thus provide the brain with signals about head orientation with respect to gravity. Such information is crucial to maintain one's body in a vertical orientation and to orient oneself in the physical world (Barra et al. 2010).

3.2 Central mechanisms

The vestibulo-thalamo-cortical pathways that transmit vestibular information from the peripheral vestibular organs to the cortex involve several structures relaying and processing vestibular sensory signals. We describe below vestibular sensory processing in the vestibular nuclei complex, thalamus, and cerebral cortex.

⁸ As we will see below, the neural signal provided by the peripheral vestibular organs does not allow us to distinguish whether the self is (active motion) or is not (passive motion) the agent of the action. Therefore, peripheral vestibular signals are ambiguous regarding the sense of agency. Yet, comparisons with motor efference copy in several vestibular neural structures allow such distinction and provide a sense of agency.

3.2.1 The vestibular nuclei complex and thalamus

The eighth cranial nerve transmits vestibular signals from the vestibular end organs to the vestibular nuclei complex and cerebellum (Barmack 2003). The vestibular nuclei complex is located in the brainstem and is the main relay station for vestibular signals. From the vestibular nuclei, descending projections to the spinal cord are responsible for vestibulo-spinal reflexes and postural control. Ascending projections to the oculomotor nuclei support eye movement control, while ascending projections to the thalamus and subsequently to the neocortex support the vestibular contribution to higher brain functions. Vestibular nuclei are also strongly interconnected with several nuclei in the brainstem and limbic structures, enabling the control of autonomic functions and emotion (see section 4.1.3) (Balaban 2004; Taube 2007).

The role of the vestibular nuclei is not limited to a relay station for vestibular signals. Complex sensory processing takes place in vestibular nuclei neurons, involving, for example, the distinction between active, self-generated head movements and passive, externally imposed head movements (Cullen et al. 2003; Roy & Cullen 2004). As we will argue in section 4.4, this processing is likely to play a crucial role in the sense of agency, especially concerning full-body agency during locomotion. Another characteristic of the vestibular nuclei complex is the large extent of *multisensory convergence* that occurs within it (Roy & Cullen 2004; Tomlinson & Robinson 1984; Waespe & Henn 1978), which leads to the perceptual “disappearance” of vestibular signals as they are merged with eye movement, visual, tactile, and proprioceptive signals. Because there is “no overt, readily recognizable, localizable, conscious sensation” from the vestibular organs during active head movements, excluding artificial passive movements and pathological rotatory vertigo, the vestibular sense has been termed a “silent sense” (Day & Fitzpatrick 2005).

Ascending projections from the vestibular nuclei complex reach the thalamus. These projections are bilateral and very distributed as

there is no thalamic nucleus specifically dedicated to vestibular processing, as compared to visual, auditory, or tactile processing.⁹ Anatomical and electrophysiological studies in rodents and primates identified vestibular neurons in many thalamic nuclei (review in Lopez & Blanke 2011). Important vestibular projections have been noted in the ventroposterior complex of the thalamus, a group of nuclei typically involved in somatosensory processing (Marlinski & McCrea 2008a; Meng et al. 2007). Other vestibular projections have been identified in the ventroanterior and ventrolateral nuclear complex, intralaminar nuclei, as well as in the lateral and medial geniculate nuclei (Kotchabhakdi et al. 1980; Lai et al. 2000; Meng et al. 2001). Electrophysiological studies revealed that similarly to vestibular nuclei neurons, thalamic vestibular neurons can distinguish active, self-generated head movements from passive head movements, showing a convergence of vestibular and motor signals in the thalamus (Marlinski & McCrea 2008b).

3.2.2 Vestibular projections to the cortex

Vestibular processing occurs in several cortical areas as demonstrated as early as the 1940s in the cat neocortex and later in the primate neocortex (reviews in Berthoz 1996; Fukushima 1997; Grüsser et al. 1994; Guldin & Grüsser 1998; Lopez & Blanke 2011). Figure 3 summarizes the main vestibular areas found in the monkey and human cerebral cortex. More than ten vestibular areas have been identified to date.

Electrophysiological and anatomical studies in animals have revealed important vestibular projections to a region covering the posterior parts of the insula and lateral sulcus, an area referred to as the parieto-insular vestibular cortex (PIVC) (Grüsser et al. 1990a; Guldin et al. 1992; Liu et al. 2011). Other vestibular regions include the primary somatosensory cortex (the hand and neck somatosensory representations of postcentral areas 2 and 3 [Ödkvist et al. 1974;

⁹ Olfactory processing in the thalamus seems also to be different from processing of the main senses as there is no direct relay between sensory neurons and primary cortex, and olfactory thalamic nuclei have been identified only recently (Courtillot & Wilson 2014).

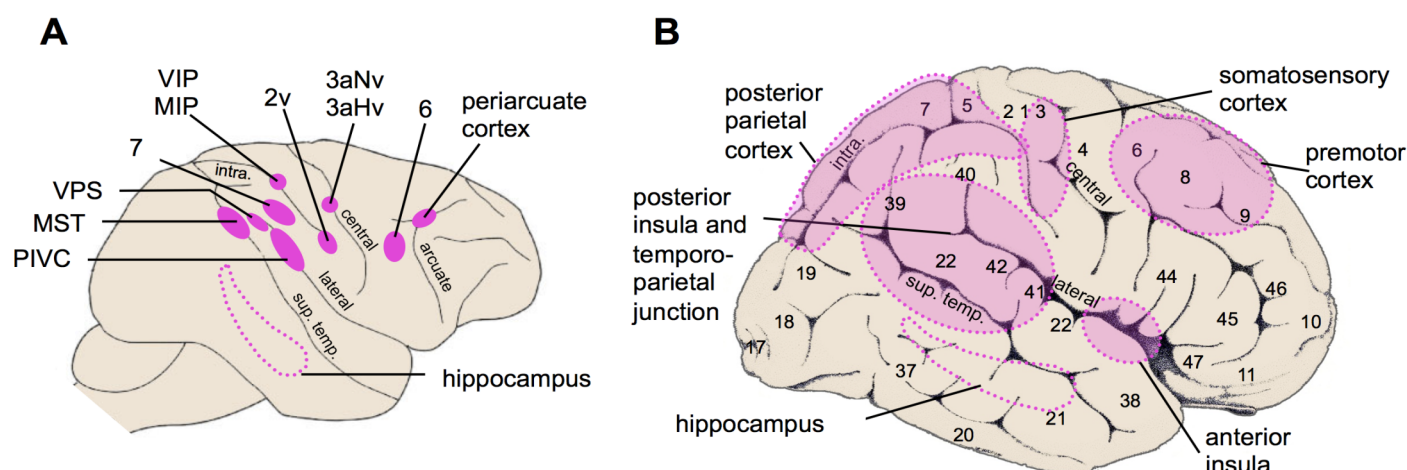


Figure 3: Schematic representation of the main cortical vestibular areas. (A) Main vestibular areas in monkeys are somatosensory areas 2v and 3av (3aHv (3a-hand-vestibular region), 3aNv (3a-neck-vestibular region)) in the postcentral gyrus, frontal area 6v and the periarculate cortex, parietal area 7, MIP (medial intraparietal area) and VIP (ventral intraparietal area), extrastriate area MST (medial superior temporal area), PIVC (parieto-insular vestibular cortex), VPS (visual posterior sylvian area), and the hippocampus. Major sulci are represented: arcuate sulcus (arcuate), central sulcus (central), lateral sulcus (lateral), intraparietal sulcus (intra.), and superior temporal sulcus (sup. temp.). Adapted from Lopez and Blanke after Sugiuchi et al. (2005). (B) Main vestibular areas in the human brain identified by noninvasive functional neuroimaging techniques. Numbers on the cortex refer to the cytoarchitectonic areas defined by Brodmann. Adapted from Lopez & Blanke (2011) after Sugiuchi et al. (2005).

Schwarz et al. 1973; Schwarz & Fredrickson 1971]); ventral and medial areas of the intraparietal sulcus (Bremmer et al. 2001; Chen et al. 2011; Schlack et al. 2005); visual motion sensitive area MST (Bremmer et al. 1999; Gu et al. 2007); frontal cortex (motor and premotor cortex and the frontal eye fields [Ebata et al. 2004; Fukushima et al. 2006]); cingulate cortex (Guldin et al. 1992) and hippocampus (O'Mara et al. 1994). These findings indicate that vestibular processing in the animal cortex relies on a highly distributed cortical network.

A similar conclusion has been drawn from neuroimaging studies conducted in humans. These studies have used fMRI and PET during caloric and galvanic vestibular stimulation¹⁰ and

revealed that the human vestibular cortex closely matches the vestibular regions found in animals. Vestibular responses were found in the insular cortex and parietal operculum as well as in several regions of the temporo-parietal junction (superior temporal gyrus, angular and supramarginal gyri). Other vestibular activations are located in the primary and secondary somatosensory cortex, precuneus, cingulate cortex, frontal cortex, and hippocampus (Bense et al. 2001; Bottini et al. 1994; Bottini et al. 1995; Dieterich et al. 2003; Eickhoff et al. 2006; Indovina et al. 2005; Lobel et al. 1998; Suzuki et al. 2001).

It is of note that the non-human animal and human vestibular cortex differs from other sensory cortices as there is apparently no *primary vestibular cortex*; that is, there is no koniocortex dedicated to vestibular processing and containing only or mainly vestibular responding neurons (Grüsser et al. 1994; Guldin et al. 1992; Guldin & Grüsser 1998), stressing again the multisensory character of the vestibular

behind one ear, and the cathode on the opposite side. The cathodal current increases the firing rate in the ipsilateral vestibular afferents.

¹⁰ Caloric and galvanic vestibular stimulations are the two most common techniques to artificially (i.e., without any head or full-body movements) stimulate the vestibular receptors. Caloric vestibular stimulation was developed by Robert Bárány and consists of irrigating the auditory canal with warm (e.g., 45°C) or cold (e.g., 20°C) water (or air), creating convective movements of the endolymphatic fluid mainly in the horizontal semicircular canals. This stimulation evokes a vestibular signal close to that produced during head rotations. Galvanic vestibular stimulation consists of the application of a transcutaneous electrical current through electrodes placed on the skin over the mastoid processes (i.e., behind the ears). Galvanic vestibular stimulation is often applied binaurally, with the anode fixed

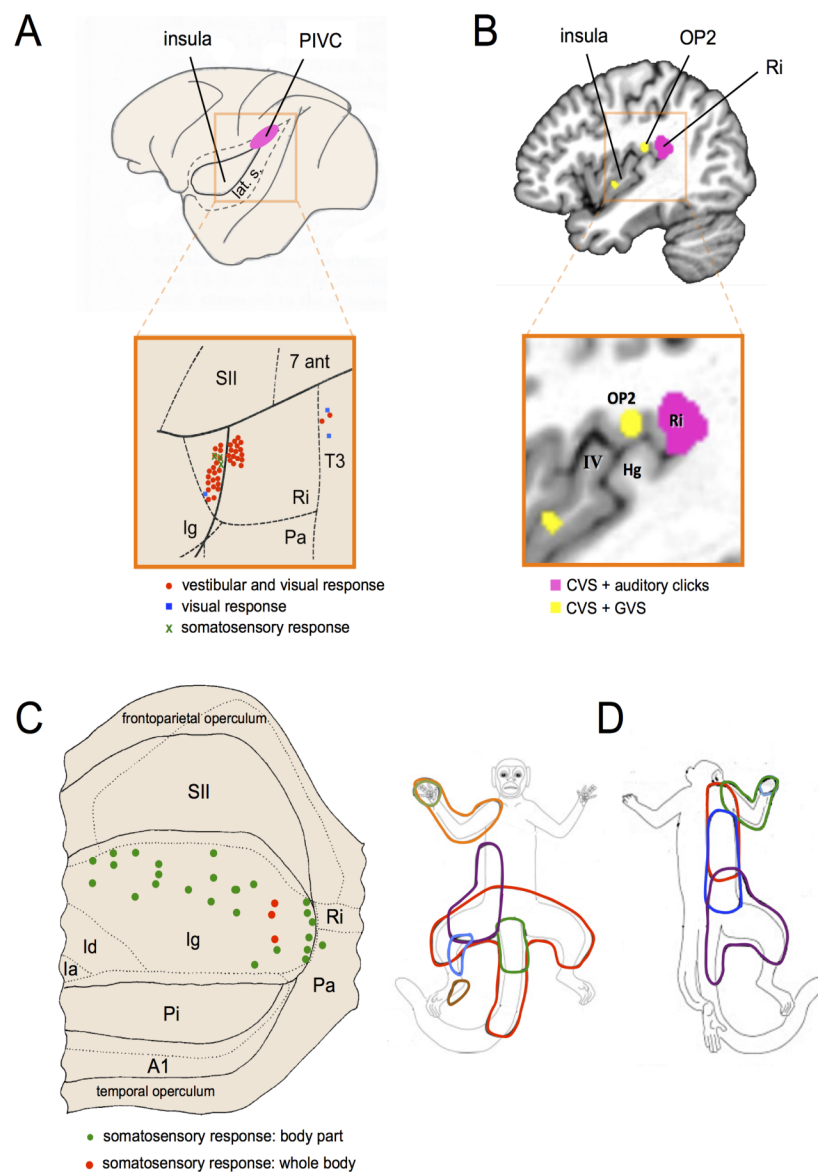


Figure 4: Anatomical location and functional properties of the parieto-insular vestibular cortex (PIVC). (A) Schematic representation of the macaque brain showing the location of the PIVC. For the purpose of illustration, the lateral sulcus (lat. s.) is shown unfolded. The macaque PIVC is located in the parietal operculum at the posterior end of the insula and retroinsular cortex. Modified from Grüsser et al. (1994). The insert illustrates the location of vestibular neurons in different regions of the lateral sulcus in a squirrel monkey (*Saimiri sciureus*). The lateral sulcus is shown unfolded to visualize the retroinsular cortex (Ri), secondary somatosensory cortex (SII), granular insular cortex (Ig), and auditory cortex (PA). Vestibular neurons (red dots) were mostly located in Ri and Ig. Adapted from Guldin et al. (1992). (B) Vestibular activations found in the human PIVC using meta-analysis of functional neuroimaging data. The Ri showed a convergence of activations evoked by caloric vestibular stimulation (CVS) of the semicircular canals and auditory activation of the otolith organs (pink). The parietal operculum (OP2) and posterior insula showed a convergence of activations evoked by CVS and galvanic vestibular stimulation (GVS) of all primary vestibular afferents (yellow). Hg (Heschl's gyrus). Adapted from Lopez et al. (2012). (C) View of the unfolded lateral sulcus of the rhesus monkey (*Macaca mulatta*) showing somatosensory neurons (green dots) in the granular insula, of which some have large somatosensory receptive fields covering the whole body (red dots). Ia (agranular insular field); Id (dysgranular insular field); A1 (first auditory field); Pa (postauditory field); Pi (parainsular field). Modified from Schneider et al. (1993). (D) Representation of the size of the receptive fields of neurons recorded in somatosensory representations of the body found in the dorsal part of the insula (ventral somatosensory area) of the titi monkey (*Callicebus moloch*). Modified from Coq et al. (2004).

lar system. All areas processing vestibular signals are multimodal, integrating visual, tactile, and proprioceptive signals. The PIVC has been shown to occupy a key role in the cortical vestibular network and is the only vestibular area that is connected to all other vestibular regions described above. The PIVC also receives signals from the primary somatosensory cortex, premotor cortex, posterior parietal cortex, and the cingulate cortex (Grüsser et al. 1994; Guldin et al. 1992), and it integrates signals from personal and extrapersonal spaces. Given these characteristics, we believe that the PIVC should be importantly involved in a coherent representation of the bodily self and the body embedded in the world.

3.2.3 The PIVC as a core, multimodal, vestibular cortex

The group of Grüsser was the first to describe vestibular responses in the monkey PIVC. Vestibular neurons were located in several regions of the posterior end of the lateral sulcus “in the upper bank of the lateral sulcus around the posterior end of the insula, sometimes also within the upper posterior end of the insula [... and] more posteriorly in the retroinsular region or more anteriorly in the parietal operculum” (Grüsser et al. 1990a, pp. 543-544; Grüsser et al. 1990b; Guldin et al. 1992; Guldin & Grüsser 1998). Figure 4A illustrates the location of PIVC in the macaque brain. Recent investigations of PIVC in rhesus monkeys revealed that vestibular neurons were mostly located in the retroinsular cortex and at the junction between the secondary somatosensory cortex, retroinsular cortex, and granular insular cortex (area Ig) (Chen et al. 2010; Liu et al. 2011).

In humans, functional neuroimaging studies used caloric and galvanic vestibular stimulation and showed activations in and around the posterior insula and temporo-parietal junction (Bense et al. 2001; Bottini et al. 1994; Dieterich et al. 2003; Eickhoff et al. 2006; Lobel et al. 1998; Suzuki et al. 2001). Because these activations also extend to the superior temporal gyrus, posterior and anterior insula, and inferior parietal lobule, the exact location of human

PIVC is still debated (review in Lopez & Blanke 2011). Recent meta-analyses of vestibular activations suggest that the core vestibular cortex is in the *parietal operculum*, *retroinsular cortex*, and/or *posterior insula* (Lopez et al. 2012; zu Eulenburg et al. 2012) (figure 4B). Of note, several neuroimaging studies have also implicated the anterior insula in vestibular processing (Bense et al. 2001; Bottini et al. 2001; Fasold et al. 2002). The insula is crucial for interoceptive awareness (Craig 2009) and could provide the neural substrate for vestibulo-interoceptive interactions that impact several aspects of the bodily self (see section 4.1.3).

4 Vestibular contributions to various aspects of the bodily self

The aim of this section is to describe several mechanisms by which the vestibular system might influence multisensory mechanisms underlying the bodily self. Again, we would like to stress that the vestibular system seems of utter importance for the most minimal aspects of self-consciousness (i.e., the sense of location in a spatial reference frame) (Windt 2010; Metzinger 2013, 2014) but at the same time also contributes to our rich sense of a bodily self in daily life. We will try to include both aspects in the following section. We further point out that while some mechanisms of a vestibular contribution to the sense of a self are now accepted, others are still largely speculative. We start by pinpointing the influence of the vestibular system on basic bodily senses such as touch and pain (section 4.1, which are subjectively experienced as bodily, i.e., as coming from within one's own bodily borders, and thus importantly contribute to a sense of bodily self. We then outline evidence for a vestibular contribution to several previously identified and experimentally modified components of the multisensory bodily self: body schema and body image, body ownership, agency, and self-location (sections 4.2–4.5). On the basis of recent data on self-motion perception in a social context and on the existence of shared sensorimotor representations between one's own body and others' bodies, we propose a vestibular contribution to the socially embedded self (section 4.6).

4.1 The sensory self

4.1.1 Touch

The feeling of touch, as its subjective perception is confined within the bodily borders, is considered as crucial for the feeling of ownership and other aspects of the bodily self (Makin et al. 2008) and a loss of somatosensory signals has been associated with a disturbed sense of the bodily self (e.g., Lenggenhager et al. 2012). Vestibular processes have been shown to interact with the perception and location of tactile stimulation. Clinical studies in brain-damaged patients suffering from altered somatosensory perceptions showed transient improvement of somatosensory perception during artificial vestibular stimulation (Kerkhoff et al. 2011; Vallar et al. 1990). Furthermore, studies in healthy participants showed that caloric vestibular stimulation can alter conscious perception of touch (Ferrè et al. 2011), probably due to interfering effects in the parietal operculum (Ferrè et al. 2012). A recent study further suggests that vestibular stimulations not only modify tactile perception thresholds, but also the perceived location of stimuli applied to the skin (Ferrè et al. 2013), a finding likely related to a vestibular influence on the body schema (see section 4.2).

Behavioural evidence of vestibulo-tactile interactions is in line with both human and animal physiological and anatomical data. Human neuroimaging studies identified areas responding to tactile, proprioceptive, and caloric vestibular stimulation in the posterior insula, retroinsular cortex, and parietal operculum (Bottini et al. 1995; Bottini et al. 2001; Bottini et al. 2005; zu Eulenburg 2013). Electrophysiological recordings in monkeys revealed a *vestibulo-somesthetic convergence* in most of the PIVC neurons. Bimodal neurons in the PIVC have large somatosensory receptive fields often located in the region of the neck and respond to muscle pressure, vibrations, and rotations applied to the neck (Grüsser et al. 1990b).

To date, the influence of caloric vestibular stimulation on somatosensory perception has been measured at the level of peripheral body parts only (e.g., the capacity to detect

touch applied to the hand, or to locate touch on the hand), but not on more central body parts or *the entire body*. Here, we propose that vestibular signals are not only important for sensory processes and awareness of body parts, but even more for *full-body awareness*. This hypothesis is supported by findings from mapping of the posterior end of the lateral sulcus in rhesus monkey that revealed neurons in the granular field of the posterior insula with *large* and *bilateral* tactile receptive fields (Schneider et al. 1993). The range of stimuli used included brushing and stroking the hair, touching the skin, muscles and other deep structures, and manipulating the joints. Importantly, the authors noted that some neurons had receptive fields covering the entire surface of the animal body, excluding the face. As can be seen in figure 4C, those neurons (red dots) were located in the most posterior part of the insula. Functional mapping conducted in the dorsal part of the insula in other monkey species has also identified neurons with large and sometimes bilateral tactile receptive fields (Coq et al. 2004) (figure 4D). So far, there is no direct evidence that neurons with full-body receptive fields receive vestibular inputs, probably because to date few electrophysiological studies have directly investigated the convergence of vestibular and somatosensory signals in the lateral sulcus (Grüsser et al. 1990a; Grüsser et al. 1990b; Guldin et al. 1992). We hypothesize that caloric and galvanic vestibular stimulation, as well as physical head rotations and translations, are likely to interfere with populations of neurons with whole-body somatosensory receptive fields and therefore may strongly impact full-body awareness. Indeed, in daily life the basic sense of touch, especially regarding large body segments, should be crucial to experience a bodily self. While full-body tactile perception hasn't been directly assessed during vestibular stimulation, the fact that caloric vestibular stimulation in healthy participants as well as acute vestibular dysfunction can evoke the feeling of strangeness and numbness for the entire body might point in this direction (see Lopez 2013, for a review).

4.1.2 Pain

Similar to touch, the experience of pain has been described as crucial to self-consciousness and the feeling of an embodied self. In his book “Still Lives—Narratives of Spinal Cord Injury” (Cole 2004), the neurophysiologist Jonathan Cole reports the case of a patient with a spinal cord lesion who described that “the pain is almost comfortable. Almost my friend. I know it is there, it puts me in contact with my body” (p. 89). This citation impressively illustrates how important the experience of pain might be in some instances for the sense of a bodily self. Reciprocal relations between pain and the sense of self are further supported by observations of altered pain perception and thresholds during dissociative states of bodily self-consciousness, such as depersonalization (Röder et al. 2007), dissociative hypnosis (Patterson & Jensen 2003) and out-of-body experiences (Green 1968). Similarly, acting in an immersive virtual environment is also associated with an increase in pain thresholds (Hoffman et al. 2004), a fact that is now increasingly exploited in virtual reality based pain therapies. This increase in pain threshold depends on the strength of *feeling of presence* in the virtual environment, i.e., the sense of “being there,” located in the virtual environment (Gutiérrez-Martínez et al. 2011; see also section 4.5.2.1). These analgesic effects of immersion and presence in virtual realities are usually explained by attentional resource mechanisms (i.e., attention is directed to the virtual environment rather than the painful event). Yet, all described instances involve also illusory self-location which has shown in full-body illusions to be accompanied by an increasing in pain thresholds or altered arousal response to painful stimuli (Hänsel et al. 2011; Romano et al. 2014). We thus speculate that analgesic effects of immersion could also be linked to disintegrated multisensory signals and a related illusory change in self-location and global self-identification. Since the vestibular system is crucially involved in self-location (see section 4.5), we suggest that some interaction effects between altered self-location and pain may be

mediated by the vestibular system.¹¹ Interestingly, galvanic and caloric stimulation, which also induce illusory changes in self-location, increase pain thresholds in healthy participants (Ferrè et al. 2013). This result and several clinical observations suggest an interplay between vestibular processes, nociceptive processes, and the sense of the bodily self (André et al. 2001; Balaban 2011; Gilbert et al. 2014; McGeoch et al. 2008; Ramachandran et al. 2007).

These interactions are likely to rely on multimodal areas in the insular cortex. Intracranial electrical stimulations of the posterior insula in conscious epileptic patients revealed nociceptive representations with a somatotopic organization (Mazzola et al. 2009; Ostrowsky et al. 2002). Functional neuroimaging studies in healthy participants also demonstrated that painful stimuli (usually applied to the hand or foot) activate the operculo-insular complex (Baumgartner et al. 2010; Craig 2009; Kurth et al. 2010; Mazzola et al. 2012; zu Eulenburg et al. 2013). It has to be noted that vestibulo-somesthetic convergence may also exist in thalamic nuclei such as the ventroposterior lateral nucleus, known to receive both somatosensory and vestibular signals (Lopez & Blanke 2011). The parabrachial nucleus of the brainstem is also a region where vestibular and nociceptive signals converge, as shown by noxious mechanical and thermal cutaneous stimulations (Balaban 2004; Bester et al. 1995). The parabrachial nucleus is further strongly interconnected with the insula and amygdala and may control some autonomic manifestations of pain (Herbert et al. 1990). Furthermore, a recent fMRI study revealed an overlap between brain activations caused by painful stimuli and by artificial vestibular stimulation in the anterior insula (zu Eulenburg et al. 2013), a structure that has been proposed to link the homeostatic evaluation of the current state of the bodily self to broader social and motivational aspects (Craig 2009). We speculate that such association could explain why illusory changes in

¹¹ A recent study investigating pain thresholds during the rubber hand illusion did not show any change in pain threshold or perception (Mohan et al. 2012), suggesting that pain perception is linked more to global aspects of the bodily self, e.g., self-location.

self-location during vestibular stimulation or during full-body illusions decrease pain thresholds.

4.1.3 Interoception

Visceral signals and their cortical representation—often referred to as *interoception*—are thought to play a core role in giving rise to a sense of self (e.g., [Seth 2013](#)). It has been proposed that visceral signals influence various aspects of emotional and cognitive processes (e.g., [Furman et al. 2013](#); [Lenggenhager et al. 2013](#); [Werner et al. 2014](#); [van Elk et al. 2014](#)) and anchor the self to the physical body ([Maister & Tsakiris 2014](#); [Tsakiris et al. 2011](#)). For this reason, various clinical conditions involving disturbed self-representation and dissociative states have been related to abnormal interoceptive processing ([Seth 2013](#), but see also [Michal et al. 2014](#) for an exception). Further evidence that interactions of exteroceptive with interoceptive signals play a role in building a self-representation comes again from research using bodily illusions in healthy participants. Two recent studies introduced an interoceptive version of the rubber hand illusion ([Suzuki et al. 2013](#)) and the full-body illusion ([Aspell et al. 2013](#)). In both cases, a visual cue on the body part/full body was presented in synchrony/asynchrony with the participant's own heartbeat. Synchrony increased self-identification with the virtual hand or body and modified the experience of self-location, thus suggesting a modulation of these components through interoceptive signals.

Vestibular processing in the context of such interoceptive bodily illusions has not yet been studied. Yet, we would like to emphasize the important interactions between the vestibular system and the regulation of visceral and autonomic functions at both functional and neuroanatomical levels (review in [Balaban 1999](#)). As mentioned earlier, the coding of body orientation in space relies on otolithic information signaling the head orientation with respect to gravity. Self-orientation with respect to gravity also requires that the brain integrates these vestibular signals with information

from gravity receptors in the trunk (e.g., visceral signals from kidneys and blood vessels) ([Mittelstaedt 1992](#); [Mittelstaedt 1996](#); [Vaitl et al. 2002](#)). Other examples of interactions between the vestibular system and autonomic regulation come from the vestibular control of blood pressure, heart rate, and respiration ([Balaban 1999](#); [Jauregui-Renaud et al. 2005](#); [Yates & Bronstein 2005](#)). Blood pressure, for instance, needs to be adapted as a function of body position in space and the vestibular signals are crucially used to regulate the baroreflex. Vestibular-mediated symptoms of motion sickness such as pallor, sweating, nausea, salivation, and vomiting are also very well-known and striking examples of the vestibular influence on autonomic functions.

At the anatomical level, there is a large body of data showing that vestibular information projects to several brain structures involved in autonomic regulation, including the *parabrachial nucleus*, nucleus of the solitary tract, paraventricular nucleus of the hypothalamus, and the central nucleus of the amygdala. Important research has been conducted in the monkey and rat parabrachial nucleus as this nucleus contains neurons responding to natural vestibular stimulation ([McCandless & Balaban 2010](#)) and is involved in the ascending pain pathways and cardiovascular pathways to the cortex and amygdala ([Bester et al. 1995](#); [Feil & Herbert 1995](#); [Herbert et al. 1990](#); [Jasmin et al. 1997](#); [Moga et al. 1990](#)). The parabrachial nucleus receives projections from several cortical regions, including the insula, as well as from the hypothalamus and amygdala ([Herbert et al. 1990](#); [Moga et al. 1990](#)). Accordingly, the parabrachial nucleus should be a crucial brainstem structure for basic aspects of the self as it is a place of convergence for nociceptive, visceral, and vestibular signals.

While research on the effects of vestibular stimulation on interoceptive awareness is still missing, we propose that artificial vestibular stimulation might be a particularly interesting means to manipulate interoception and investigate its influence on the sense of a bodily self.

4.2 Body schema and body image

Here, we propose that vestibular signals are not only important for the interpretation of basic somatosensory (tactile, nociceptive, interoceptive) processes, but as a consequence also contribute to *body schema* and *body image*. Body schema and body image are different types of models of motor configurations and body metric properties, including the size and shape of body segments (e.g., [Gallagher 2005](#); [de Vignemont 2010](#); [Berlucchi & Aglioti 2010](#); [Longo & Hag-gard 2010](#)). Although body schema and body image are traditionally thought to be of mostly proprioceptive and visual origin, respectively, a vestibular contribution was already postulated over a century ago (review in [Lopez 2013](#)). [Pierre Bonnier \(1905\)](#) described several cases of distorted bodily perceptions in vestibular patients and coined the term “*aschématie*” (meaning a “loss” of the *schema*) to describe these distorted perceptions of the volume, shape, and position of the body. [Paul Schilder \(1935\)](#) also noted distorted body schema and image in vestibular patients claiming for example that their “neck swells during dizziness,” “extremities had become larger,” or “feet seem to elongate.” The contribution of vestibular signals to mental body representations has been recognized more recently by Jacques Paillard. He proposed that “the ubiquitous geotropic constraint [i.e., gravitational acceleration, which is detected and coded by vestibular receptors] dominates the [body-, world-, object- and retina-centered] reference frames that are used in the visuomotor control of actions and perceptions, and thereby becomes a crucial factor in linking them together” ([Paillard 1991](#), p. 472). According to Paillard, gravity signals would help merge and give coherence to the various reference frames underpinning action and perception.

Because humans have evolved under a constant gravitational field, human body representations are strongly shaped by this physical constraint. In particular, grasping and reaching movements are constrained by gravito-inertial forces and internal models of gravity ([Indovina et al. 2005](#); [Lacquaniti et al. 2013](#); [McIntyre et al. 2001](#)). Thus, the body schema and action

potentialities must take into account signals from the otolithic sensors. For example, when a subject is instructed to reach a target while his entire body is rotated on a chair, the body rotation generates Coriolis and centrifugal forces deviating the hand. Behavioural studies demonstrate that vestibular signals generated during whole-body rotations are used to correct the hand trajectory ([Guillaud et al. 2011](#)). Other studies demonstrate that vestibular signals continuously update the body schema during hand actions. [Bresciani et al. \(2002\)](#) asked participants to point to previously memorized targets located in front of them ([figure 5A](#)). At the same time, participants received bilateral galvanic vestibular stimulation, with the anode on one side and the cathode on the other side. The data indicate that the hand was systematically deviated toward the side of anodal stimulation ([figure 5B](#)). It is important to note that galvanic vestibular stimulation is known to evoke illusory body displacements in the frontal plane and thus modifies the perceived self-location ([Fitzpatrick et al. 2002](#); see also section 4.5). One possible interpretation of the change in hand trajectory during the pointing movement was that it compensated for an “apparent change in the spatial relationship between the target and the hand,” evoked by the vestibular stimulation ([Bresciani et al. 2002](#)). Thus, vestibular signals are used to control the way we act and interact with objects in the environment.

After having established the contribution of vestibular signals to hand location and motion, we shall describe the role of vestibular signals in the perception of the body’s metric properties (the perceived shape and size of body segments). During parabolic flights, known to create temporary weightlessness and thus mimic a deafferentation of the otolithic vestibular sensors, [Lackner \(1992\)](#) reported cases of participants experiencing a “telescoping motion of the feet down and the head up internally through the body,” that is, an inversion of their body orientation. Experiments conducted on animals born and raised in hypergravity confirm an influence of vestibular signals on body representations. In these animals, changes in the strength of the gravita-

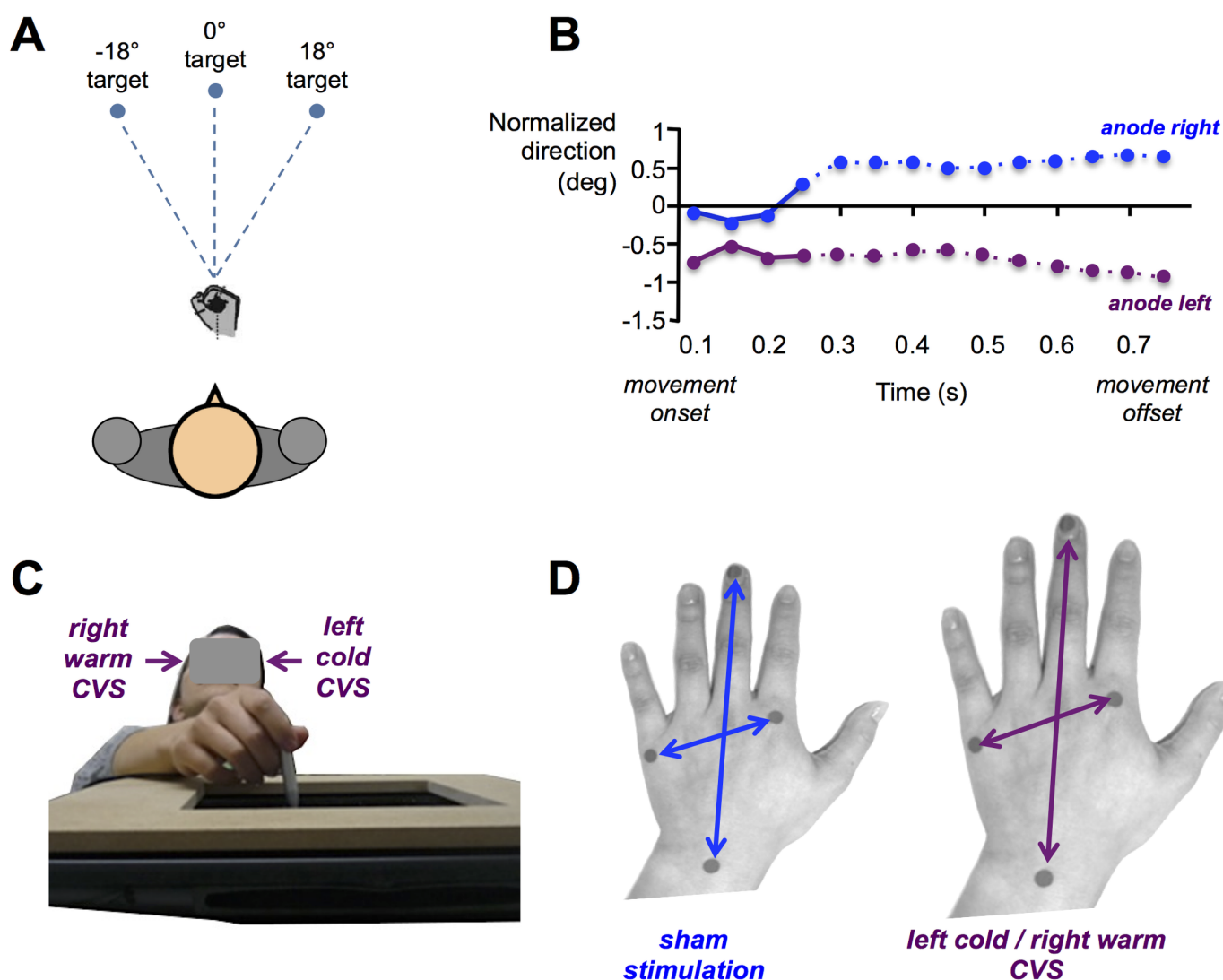


Figure 5: Influence of vestibular signals on motor control and perceived body size. (A) Pointing task toward memorized targets. Participants received binaural galvanic vestibular stimulation as soon as they initiated the hand movement (with eyes closed). (B) Deviation of the hand trajectory towards the anode (modified after [Bresciani et al. 2002](#)). (C) Proprioceptive judgment task used to estimate the perceived size of the left hand. Participants were tested blindfolded and used a stylus held in their right hand to localize on a digitizing tablet four anatomical landmarks corresponding to the left hand under the tablet. (D) Illustration of the perception of an enlarged hand during caloric stimulation activating the right cerebral hemisphere (modified after [Lopez et al. 2012](#)).

tional field permanently disorganized the somatosensory maps recorded in their primary somatosensory cortex ([Zennou-Azogui et al. 2011](#)).

Experimental evidence of a vestibular contribution to the coding of body metric properties comes from the application of stimulation in healthy participants. In a recent study, [Lopez et al. \(2012\)](#) showed that caloric vestibular stimulation modified the perceived size of the body during a proprioceptive judg-

ment task ([figure 5C](#)). Participants had their left hand palm down on a table. Above the left hand, there was a digitizing tablet on which participants were instructed to localize four anatomical targets enabling the calculation of the perceived width and length of the left hand. While participants pointed repeatedly to these targets, they received bilateral caloric vestibular stimulation known to stimulate the right cerebral hemisphere in which the left hand is mostly represented

(e.g., warm air in the right ear and cold air in the left ear). The results showed that in comparison to a control stimulation (injection of air at 37°C in both ears), in the stimulation condition the left hand appeared significantly enlarged (figure 5D), showing that vestibular signals can modulate internal models of the body.

4.3 Body ownership

Correct self-attribution of body parts and self-identification with the entire body relies on successful integration of multisensory information as evidenced by various bodily illusions in healthy participants (e.g., Botvinick & Cohen 1998; Lenggenhager et al. 2007; Petkova & Ehrsson 2008). So far there is only little evidence of a vestibular contribution to the sense of body ownership. Bisiach et al. (1991) described a patient with a lesion of the right parieto-temporal cortex who suffered from somatoparaphrenia, claiming that her left hand did not belong to her. In this patient, caloric vestibular stimulation transiently restored normal ownership for her left hand. Similarly, Lopez et al. (2010) applied galvanic vestibular stimulation to participants experiencing the rubber hand illusion and showed that the vestibular stimulation increased the feeling of ownership for the fake hand. The authors have linked such interaction between the vestibular system, multisensory integration, and body ownership to overlapping cortical areas in temporo-parietal areas and the posterior insula. No study has so far investigated the effect of vestibular stimulation on full-body ownership. Yet, reports from patients with acute vestibular disturbances as well as reports from healthy participants during caloric vestibular stimulation (Lopez 2013; Sang et al. 2006) suggest that full-body ownership might also be modified by artificial vestibular stimulation or vestibular dysfunctions. Given the importance of the vestibular system in more global aspects of the bodily self, we predict that vestibular stimulation would influence ownership even stronger in a full-body illusion than in a body-part illusion set-up.

4.4 The acting self: Sense of agency

As mentioned earlier, the sense of being the agent of one's own actions is another crucial aspect of the sense of self. Agency relies on sensorimotor mechanisms comparing the motor efference copy with the sensory feedback from the movement, and on other cognitive mechanisms such as the expectation of a self-generated movement (Cullen 2012; Jeannerod 2003, 2006). While no study so far has directly investigated vestibular mechanisms of the sense of agency, recent progress in this direction has been made in a study investigating full-body agency during a goal-directed locomotion task (Kannape et al. 2010). Participants walked toward a target and observed their motion-tracked walking patterns applied to a virtual body projected on a large screen in front of them. Various angular biases were introduced between their real locomotor trajectory and that projected on the screen. Comparable to the classical experiments assessing agency for a body part (Fournieret & Jeannerod 1998), these authors investigated the discrepancy up to which the motion of the avatar showed on the screen was still perceived as their own. During this task, the brain does not only detect visuo-motor coherence but also vestibulo-visual coherence, and self-attribution of the seen movements is thus likely to depend on vestibular signal processing. In the following we present an example of neural coding underlying an aspect of the sense of agency in several structures of the vestibulo-thalamo-cortical pathways.

In the vestibular system, peripheral organs encode in a similar way head motions for which the subject is or is not the agent.¹² Thus, vestibular organs generate similar signals during an active rotation of the head (i.e., the person is the agent of the action) or during a passive, externally imposed, rotation of the head (i.e., the person is passively moved while sitting on a rotating chair). It is important for the central

¹² Although the coding of movements by the peripheral vestibular organs is ambiguous regarding the sense of agency, the coding is not ambiguous regarding the sense of ownership for the movements and self-other distinction. Indeed, because vestibular sensors are inertial sensors, vestibular signals are necessarily related to one's own motion and are the basis of the perception that I have (been) moved, irrespective of whether the "self" is or is not the agent of this movement.

nervous system to establish whether afferent vestibular signals are generated by active or passive head movements, and this is done at various levels. Electrophysiological studies conducted in monkeys have revealed that some vestibular nuclei neurons were silent, or had a strongly reduced firing rate, during active head rotations, whereas their firing rate was significantly modulated by passive head rotations. This indicates that vestibular signals generated by active head rotations were suppressed or attenuated. This suppression of neural responses was found in the vestibular nuclei complex (Cullen 2011; Roy & Cullen 2004), thalamus (Marlinski & McCrea 2008b) and cerebral cortex, for example in areas of the intraparietal sulcus (Klam & Graf 2003, 2006). Several studies were conducted to determine which signal might induce such suppression. Roy & Cullen (2004) suggested that a *motor efference copy* was used. They showed that the suppression occurred “only in conditions in which the activation of neck proprioceptors matched that expected on the basis of the neck motor command”, suggesting that “vestibular signals that arise from self-generated head movements are inhibited by a mechanism that compares the internal prediction of the sensory consequences by the brain to the actual resultant sensory feedback” (p. 2102). In conclusion, as early as the first relay along the vestibulo-thalamo-cortical pathways, neural mechanisms have the capacity to distinguish between the consequences of active and passive movements on vestibular sensors. Given this evidence, we suggest an important contribution of the vestibular system to the sense of agency in general and to full-body agency in particular.

4.5 The spatial self: Self-location

4.5.1 Behavioural studies in humans

Self-location is the experience of where “I” am located in space and is one of the (if not the) crucial aspects of the bodily self (Blanke 2012). Recently, self-location has been systematically investigated in human behavioural and neuroimaging studies using multisensory conflicts (Ionta et al. 2011; Lenggenhager et al.

2007; Lenggenhager et al. 2009; Pfeiffer et al. 2013). While we usually experience ourselves as located within our own bodily borders at one single location in space, the sense of self-location can be profoundly disturbed in psychiatric and neurological conditions, most prominently during *out-of-body experiences* (Bunning & Blanke 2005). Based on findings in neurological patients that revealed a frequent association between vestibular illusions (floating in the room, sensation of lightness or levitation) and out-of-body experiences, Blanke and colleagues proposed that the illusory disembodied self-location was due to a dis-integration of vestibular signals with signals from the personal (tactile and proprioceptive signals) and extrapersonal (visual) space (Blanke et al. 2004; Blanke & Mohr 2005; Blanke 2012; Lopez et al. 2008). The authors proposed that this multisensory disintegration is mostly a result of abnormal neural activity in the temporo-parietal junction (Blanke et al. 2005; Blanke et al. 2002; Heydrich & Blanke 2013; Ionta et al. 2011). In this section, we review experimental data in healthy participants that may account for the tight link between vestibular disorders and illusory or simulated changes in self-location. While the most direct evidence of such a link comes from the finding that artificial stimulation of the vestibular organs induces an illusory change in self-location¹³ (Fitzpatrick & Day 2004; Fitzpatrick et al. 2002; Lenggenhager et al. 2008), we focus on three experimental set-ups that have been used to alter the experience of self-location in healthy participants.

4.5.1.1 Illusory change in self-location during full-body illusions

Full-body illusions have increasingly been used to study the mechanisms underlying self-location (see Blanke 2012, for a review). No study has so far investigated the influence of artificial vestibular stimulation on such illusions. Nevertheless, there is some experimental evidence suggesting a vestibular involvement in illusory changes in self-location. While the initial full-

¹³ Depending on the stimulation parameters and method, participants describe various sensations of movements and change in position.

body illusion was described in a standing position (Lenggenhager et al. 2007), the paradigm has later been adapted to a lying position (Ionta et al. 2011; Lenggenhager et al. 2009; Pfeiffer et al. 2013), mainly because the frequency of spontaneous out-of-body experiences is higher in lying position than in standing or sitting positions (Green 1968). It has been speculated that this influence of the body position on the sense of embodiment is related to the decreased sensitivity of otolithic vestibular receptors and decreased motor and somatosensory signals in the lying position (Pfeiffer et al. 2013). We hypothesized that under such conditions of reduced vestibular (and proprioceptive) information, visual capture is enhanced in situations of multisensory conflict, thus resulting in a stronger change in self-location during the full-body illusion. So far, the full-body illusion has not been directly compared in standing versus lying positions. However, the application of visuo-tactile conflicts in a lying position not only alters self-location but also evokes sensations of floating (Ionta et al. 2011; Lenggenhager et al. 2007). This finding hints toward a reweighting of visual, tactile, proprioceptive, and vestibular information during the illusion, plausibly in the temporo-parietal junction and human PIVC. In line with this finding, the changes in self-location and perspective have been associated with individual perceptual styles of visual-field dependence (Pfeiffer et al. 2013), i.e., weighting of visual as compared to vestibular information in a subjective visual vertical task, suggesting an individually different contribution and weighting of the various senses for the construction of the bodily self (for a similar finding regarding the rubber hand illusion, see David et al. 2014).

4.5.1.2 Mental own-body transformation and perspective taking

Another way to investigate bodily self-consciousness has been to use experimental paradigms requiring participants to put themselves “into the shoes” of another individual, that is to mentally simulate an external self-location (own-body, egocentric, mental trans-

formation tasks) and a third-person visuo-spatial perspective. Typically, participants are instructed to make left-right judgments about a body, for example, to judge whether this other shown person is wearing a glove on his right or left hand (Blanke et al. 2005; Lenggenhager et al. 2008; Parsons 1987; Schwabe et al. 2009). Other tasks require that participants adopt the visual perspective of another person to decide whether a visual object is to the right or left of the other person (David et al. 2006; Lambrey et al. 2012; Vogeley & Fink 2003). Early studies have shown that the time needed for own-body mental transformations correlates with the distance or angle between the participant’s position in the physical space and the position to be simulated (Parsons 1987). It is largely admitted that own-body mental transformation is an “embodied” mental simulation that can be influenced by various sensorimotor signals from the body (e.g., Kessler & Thomson 2010). In line with this view, various experiments demonstrated that the actual body position influences mental own-body transformation of body parts (e.g., Ionta et al. 2012). Importantly, next to proprioceptive and motor mechanisms, visuo-spatial perspective taking and own-body mental transformation also require the integration of vestibular information (active or passive body motion). Thus, while most of this research looked at how body parts’ posture (e.g., of the hand) influences mental own-body (part) transformation, some recent research investigated how mental own-body transformation is influenced by vestibular cues (Candidi et al. 2013; Dilda et al. 2012; Falconer & Mast 2012; Lenggenhager et al. 2008; van Elk & Blanke 2014). All these studies revealed that vestibular signals influence mental (full) own-body transformation, confirming again the influence of the vestibular system in the sense of self-location and perspective taking.¹⁴

¹⁴ Visuo-spatial perspective-taking has not only been used in the field of spatial cognition but also in the field of social cognition. Perspective taking is a very crucial aspect of human cognition, which allows us to understand other people’s actions and emotions. The fact that the vestibular system is importantly involved in such simulations might further suggest that the vestibular system is important for social cognition (see also section 5 and Deroualle & Lopez 2014).

4.5.1.3 Change in self-location and the feeling of presence

The development of immersive virtual environments has launched a powerful research area where the mechanisms of self-location can be investigated and manipulated by the *feeling of presence*. The term “presence” stems from virtual reality technologies and commonly refers to the feeling of being immersed (“being there”) in the virtual environment. Yet, it has been argued that “presence” also reflects a more general and basic state of consciousness (Riva et al. 2011). The study of presence has thus been suggested to provide useful tools to study (self-)consciousness, with the advantage of precise experimental control (Sanchez-Vives & Slater 2005).

Similar to previously mentioned full-body illusions, a participant who is immersed in a virtual environment receives contradicting multisensory information about his or her self-location: while visual information suggests that s/he is located in a virtual world, proprioceptive information suggests that s/he is located in the real world, for example, by indicating a different body position between the physical body and the avatar. Furthermore, and contrary to the full-body illusion, the visual information often indicates that the participant is moving, whereas the proprioceptive and vestibular information suggests that he or she is sitting still. The compelling feeling of presence in virtual environments indicates that participants rely strongly on visual cues. Of note, some authors have proposed that a sort of bi-location is possible in such a situation, by which one feels to a certain degree being localized simultaneously in both the real and virtual environments (Furlanetto et al. 2013; Wissmath et al. 2011), which has also been described in a clinical condition called heautoscopy (e.g., Blanke & Mohr 2005; Brugger et al. 1994).

Neuroimaging studies in healthy participants showed that self-identification with—and self-localization at—a position of a virtual avatar seen from a third-person perspective activates the left inferior parietal lobe (Corradi-Dell’acqua et al. 2008; Ganesh et al. 2012). Corroboratively, people who are addicted to video-

games show altered processing in a left posterior area of the middle temporal gyrus (Kim et al. 2012). These studies converge in their conclusion that multimodal areas in the temporo-parietal junction are involved in altered self-localization in virtual reality. As mentioned before, the temporo-parietal junction is a main region for vestibular processing. We thus hypothesize that the feeling of presence might be mediated by vestibular signals, which should be directly tested by assessing whether the feeling of presence can be modified by caloric and galvanic vestibular stimulation.

4.5.2 Physiological and vestibular mechanisms of self-location

4.5.2.1 Categories of cells coding self-location and self-orientation

Electrophysiological investigations in rodents have identified three categories of neurons encoding specifically where the animal is located, how its head is oriented, and how the animal moves in its environment (see Barry & Burgess 2014, for a recent review). These neurons are referred to in the literature as “place cells,” “head-direction cells,” and “grid cells”. In rats, *place cells* have been recorded as early as the 1970s in the hippocampus, and later in the subiculum and entorhinal cortex (O’Keefe & Conway 1978; O’Keefe & Dostrovsky 1971; Poucet et al. 2003). The firing rate of these neurons increases when the animal is located at a specific position within the environment. This activity is strongly modulated by allocentric signals (visual references in the environment) and vestibular signals (Wiener et al. 2002). Place cells have later been identified in several other animal species including mice (McHugh et al. 1996), bats (Ulanovsky & Moss 2007), monkeys (Furuya et al. 2014; Ludvig et al. 2004; Matsumura et al. 1999; Ono et al. 1993) and humans (Ekstrom et al. 2003; Miller et al. 2013). *Head-direction cells* were first recorded in the rat postsubiculum and later in several nuclei constituting the Papez circuit, such as the dorsal thalamic nucleus and lateral mammillary nuclei (Taube 2007). They were also found in

the retrosplenial and entorhinal cortex. Electrophysiological recordings revealed that head-direction cells “discharge allocentrically as a function of the animal’s directional heading, independent of the animal’s location and ongoing behavior” (Taube 2007). Head-direction cells have also been identified in the monkey hippocampus (Robertson et al. 1999). Finally, *grid cells* have been identified in the rat medial entorhinal cortex, but also in the pre- and parasubiculum (Boccaro et al. 2010; Sargolini et al. 2006). Grid cells fire for multiple locations of the animal within its environment. Altogether, these locations form a periodic pattern, or “grid,” spanning the entire surface of the environment. More recently, electrophysiological recordings have shown grid cells in mice (Fyhn et al. 2008), bats (Yartsev et al. 2011) and monkeys (Killian et al. 2012), and even probable homologues of grid cells in the human hippocampus (Doeller et al. 2010; Jacobs et al. 2013).

4.5.2.2 Place cells in the human hippocampus and “virtual” self-location

We can only speculate about the neural mechanisms of place and head-direction specific coding in the human brain. With the non-invasive neuroimaging techniques available to date (fMRI, PET, scalp electroencephalography (EEG), near-infrared spectroscopy (NIRS)), it remains difficult to investigate neural activity of potential human homologues of place cells, head-direction cells and grid cells (for fMRI identification of grid cells, see Doeller et al. 2010). Single-unit recordings can only be achieved during rather rare intracranial EEG carried out for presurgical evaluations of drug refractory epilepsy.

In a seminal intracranial EEG study conducted in 7 epileptic patients, Ekstrom et al. (2003) identified neurons with place selectivity in the *hippocampus*. Patients were immersed in a virtual environment and played a taxi driver computer game, picking up customers at one location in the virtual town and delivering them to another location of the town. As illustrated in figure 6, a neuron recorded in the right hip-

pocampus had a significantly stronger firing rate when the patient was virtually “located” in the upper left corner than in any other location of the virtual town, showing its place selectivity. The authors found that 24% of neurons recorded in the hippocampus displayed a pattern of place selectivity, a proportion that was significantly larger than in the other brain structures they explored. Using a very similar procedure in a virtual environment in patients with intracranial electrodes, a recent study identified probable grid-like cells in humans (Jacobs et al. 2013). They were predominantly located in the entorhinal cortex and anterior cingulate cortex.

Interestingly, in both studies, patients did not physically move but moved virtually using button presses on a keyboard or a joystick. Nevertheless, the firing rate of these neurons changed as a function of the “virtual” location of the participants within the virtual environment. This observation indicates that both hippocampal “place cells” as well as entorhinal and cingulate “grid-cells” were coding the patient’s location in the virtual world on the basis of allocentric visual signals, rather than the patient’s position in the real world. Although the findings about these properties of the hippocampus have been mostly interpreted in the research field of spatial navigation and memory (Burgess & O’Keefe 2003), we make a new proposition that they can also shed light on the neural underpinnings of bodily self-consciousness, especially on how the brain localizes the self both in everyday life as well as in situations of multisensory conflicts.

As mentioned earlier, the experience of self-location can be manipulated by creating conflicts between visual cues about the location of one’s own body (or an avatar) in the external world and tactile or other somatosensory signals (Ehrsson 2007; Lenggenhager et al. 2011; Lenggenhager et al. 2009; Lenggenhager et al. 2007). These visuo-tactile conflicts can induce the perception of being located closer to the avatar. The recent use of these visual-tactile conflicts during fMRI recordings showed that the apparent changes in self-location and visuo-spatial perspective were related to signal changes in the temporo-parietal junction, not in the hippocampus (Ionta et al. 2011). It is not clear whether

hippocampal place cells' activity can be recorded with the large-scale, non-invasive functional neuroimaging techniques available. Yet, we predict that visuo-tactile conflicts, by modifying the experienced self-location, should also modify the neural activity of place cells and grid cells and their vestibular modulation (see next section), as showed during navigation in immersive virtual environments (Ekstrom et al. 2003; Jacobs et al. 2013). Future research using intracranial EEG recordings in epileptic patients should endeavour to study directly the relation between place cell activity and the experience of human self-location in situations of conflicting multisensory information.

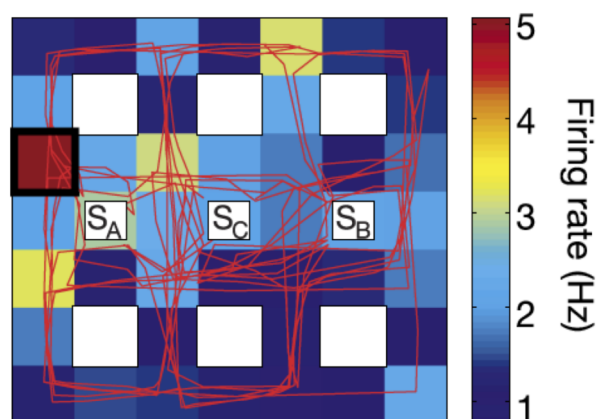


Figure 6: Map illustrating the firing rate of one cell in the right hippocampal showing a pattern of place selectivity. The rectangular map represents the virtual town explored by the participant using key presses on a keyboard and the red line represents the participant's trajectory within the virtual town. The nine white boxes indicate the location of buildings in the virtual town (SA, SB, and SC represent three shops that were "visited" by the participant). Colors from blue to red in the background represent the firing rate of the hippocampal cell as a function of the participant's location in the virtual town. This neuron displays a significantly higher firing rate when the participant was located in the left upper part of the virtual environment (location showed by a black square). Reproduced from Ekstrom et al. (2003).

4.5.2.3 Vestibular signals and place cells

In this section, we emphasize the contribution of vestibular signals to the neural coding of self-

location in the hippocampus. As mentioned above, the firing rate of place cells is strongly modulated by allocentric signals, a finding replicated in several studies in rodents (Wiener et al. 2002). Vestibular signals have also been shown to modulate the firing pattern of the hippocampal place cells, which is necessary when animals navigate in darkness (O'Mara et al. 1994).

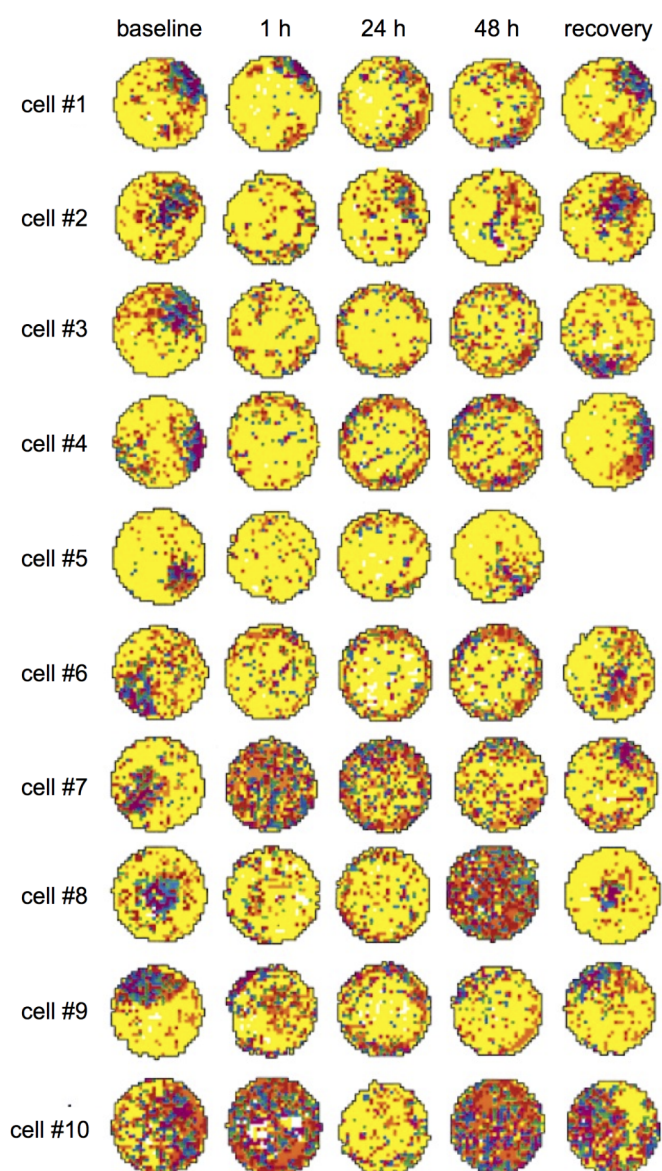


Figure 7: Modification of spatial selectivity of ten hippocampal place cells before and after inactivation of the vestibular apparatus with TTX injection. The colors ranging from yellow to purple represent the increase in firing rate of the place cells as a function of the location of the rat in the circular arena. From Stackman et al. (2002).

For example, [Stackman et al. \(2002\)](#) temporarily inactivated the vestibular system of rats using bilateral transtympanic injections of tetrodotoxin (TTX). TTX abolishes almost immediately neural activity in the vestibular nerve, producing a temporary vestibular deafferentation, mimicking the situation of patients with a bilateral vestibular loss. [Figure 7](#) illustrates changes in the firing rate of ten hippocampal neurons before and after TTX injection. Before TTX injection, hippocampal neurons displayed a typical pattern of place selectivity when the animal explored the circular environment. A major finding of this study was that as early as one hour after vestibular deafferentation, the location-specific activity of the same hippocampal neurons was strongly disturbed. In particular, the vestibular deafferentation reduced the spatial coherence and spatial information content that usually characterize the place cells. These disorders remained between thirty-six and seventy-two hours after TTX injection, despite the fact that the rats continued to explore their circular environment and had normal locomotor activity twelve hours after TTX injection. These results indicate that place cells are continuously integrating vestibular signals to estimate one's location within the physical environment and that vestibular signals strongly contribute to one of the most important neural mechanisms of self-location.

The activity of place cells or grid cells has not been recorded after vestibular deafferentation in humans. Nevertheless the neural consequences of vestibular lesions on place cells ([Stackman et al. 2002](#)) and head-direction cells ([Stackman & Taube 1997](#)) in animal models corroborate the effects of unilateral and bilateral vestibular lesions in humans. Patients with vestibular disorders may experience spatial disorientation as measured during path completion tasks ([Glasauer et al. 1994](#)) and navigation in virtual environments ([Hüfner et al. 2007](#); [Péruch et al. 1999](#)). We propose that vestibular disorders, by disorganizing the firing pattern of place cells in the human hippocampus (and in other brain regions containing place cells) may strongly disturb the sense of self-location and thus the coherent sense of self, which could

eventually even lead to disturbance of the usually very stable feeling of being located at a single place at a given time (see the strong disorganization of the place cells activity in [figure 7](#)). Another striking consequence of a bilateral vestibular loss is the induced atrophy of the hippocampus, whose volume is decreased by about seventeen percent ([Brandt et al. 2005](#)). Altogether, these data show that one neural mechanism of bodily self-location (place cells encoding of the body location in the environment) strongly relies on vestibular signals.

4.6 The socially embedded self

An important branch of research suggests that the neural mechanisms that dynamically represent multisensory bodily signals not only give rise to a sense of self, but also to the sense of others. The emerging field of social neuroscience has investigated both in animals and humans how the perception of another person modifies neural activity in body-related, sensorimotor neural processing and vice versa.¹⁵ “Sensorimotor sharing” and related mechanisms such as emotional contagion, sensorimotor resonance, or mimicry are thought to enable individuals to understand others' emotions, intentions, and actions and are thus fundamental for our social functioning. This line of research has evolved from an influential electrophysiological study that identified *mirror neurons* activated both when a monkey was performing a (body part) action and when observing someone else executing the same action ([Gallese et al. 1996](#); [Rizzolatti et al. 1996](#)). A human mirror-neuron-like system has been suggested based on neuroimaging studies that revealed similar brain activations when acting and when observing the same action being executed by another person (e.g., [Rizzolatti & Craighero 2004](#)). Importantly, similar mechanisms were found in various sensory systems as further experiments have shown common neural activity when experiencing and

¹⁵ The research on bodily illusions has recently extended to social neuroscience by investigating how sensorimotor self-other confusion (during the rubber hand, full-body, and enfacement illusions) affects the perception of another person and, vice versa, how the perception of another person influences illusory self-other confusion (e.g., [Bufalari et al. 2014](#); [Paladino et al. 2010](#); [Tajadura-Jiménez et al. 2012](#)).

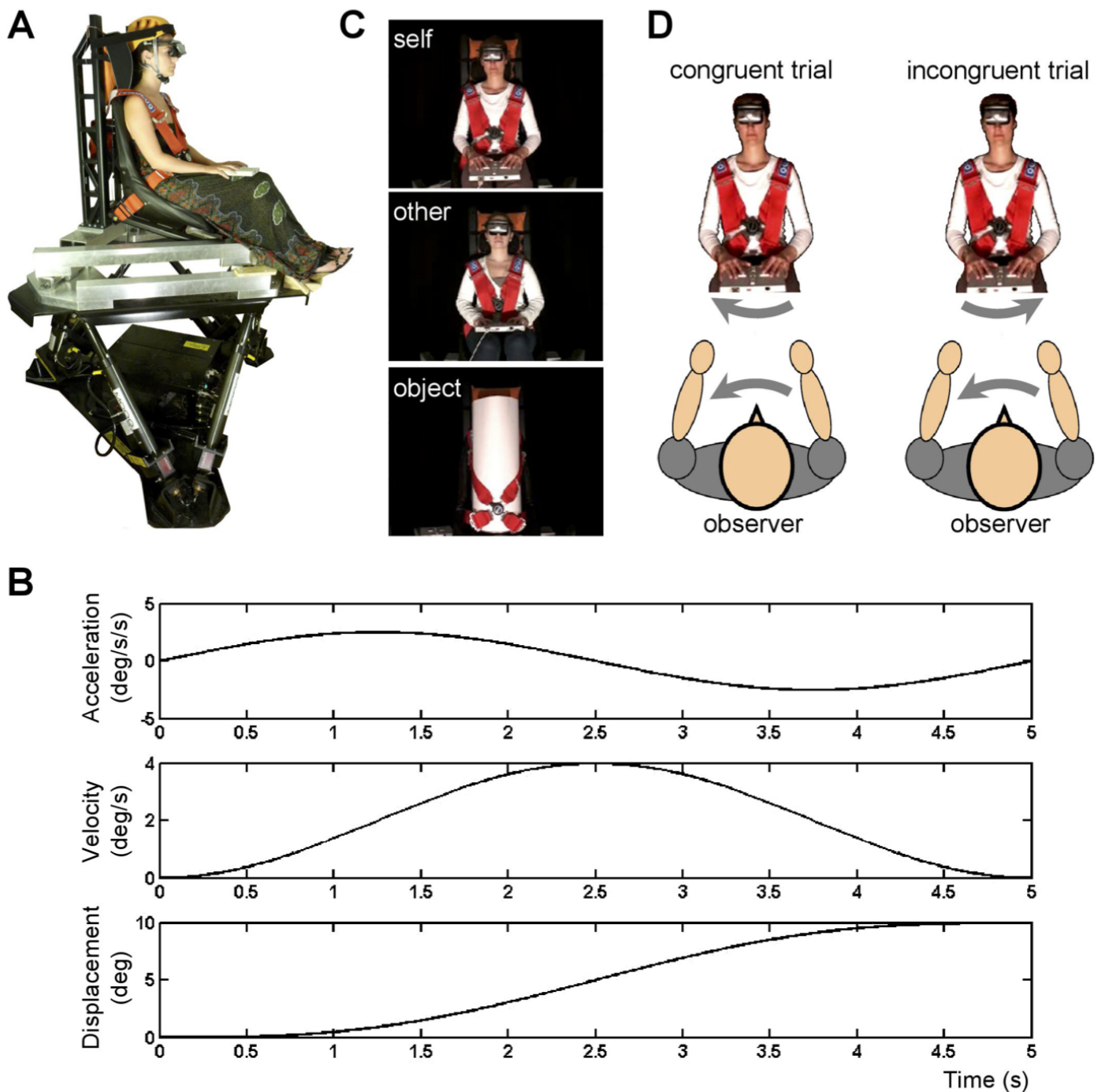


Figure 8: Experimental setup used to measure the influence of body movement observation on whole body self-motion perception. (A) Self-motion perception was tested in twenty-one observers seated on a motion platform. Motion stimuli were yaw rotations lasting for 5s with peak velocity of $0.1^\circ/\text{s}$, $0.6^\circ/\text{s}$, $1.1^\circ/\text{s}$, and $4^\circ/\text{s}$. (B) Example of a motion profile consisting of a single cycle sinusoidal acceleration. Acceleration, velocity, and displacement are illustrated for the highest velocity used at $4^\circ/\text{s}$. (C) Observers wore a head-mounted display through which 5-s videos were presented, depicting their own body, the body of another participant matched for gender and age, or an inanimate object. (D) During congruent trials, the observers and the object depicted in the video were rotated in the same direction (specular congruency). Reproduced from [Lopez et al. \(2013\)](#).

observing pain (Lamm et al. 2011, for a recent meta-analysis), when being touched and observing someone being touched (Keysers et al. 2004), and when inhaling disgusting odorants and observing the face of someone inhaling disgusting odorants (Wicker et al. 2003)

No human neuroimaging study so far has investigated brain mechanisms when experiencing a vestibular sensation and seeing somebody experiencing a vestibular sensation (e.g., being passively moved in space). Yet, recent findings from a behavioural study in humans suggest that the observation of another person's whole-body motion might influence vestibular self-motion perception (Lopez et al. 2013; see figure 8). In this study, participants were seated on a whole-body motion platform and passively rotated around their main vertical body axis. They were asked in a purely vestibular task to indicate in which direction (clockwise vs. counter-clockwise) they were rotated while looking at videos depicting their own body, another body, or an object rotating in the same plane. The spatial congruency between self-motion and the item displayed in the video was manipulated by creating congruent trials (specular congruency) and incongruent trials (non-specular congruency). The results indicated self-motion perception was influenced by the observation of videos showing passive whole-body motion. Participants were faster and more accurate when the motion depicted in the video was congruent with their own body motion. This effect depended on the agent depicted in the video, with significantly stronger congruency effects for the “self” videos than for the “other” videos, which is in line with the effects previously reported for the tactile system (Serino et al. 2009; Serino et al. 2008). Lopez et al. (2013) speculated on the existence of a *vestibular mirror neuron system* in the human brain, that is a set of brain regions activated both by vestibular signals and by observing bodies being displaced. As noted earlier, vestibular regions show important patterns of visuo-vestibular convergence in the parietal cortex, which could underlie such effects (Bremmer et al. 2002; Grüsser et al. 1990b).

On the basis of these findings as well as the data presented above on the importance of vestibular processes in spatial, cognitive, and social perspective-taking, we propose that the vestibular system is not only involved in shaping and building the perception of a bodily self but is also involved in better understanding and predicting another person's (full-body) action through sensorimotor resonance (see also Deroualle & Lopez 2014).

5 General conclusion

During the last years, various theories from psychological, neuroscientific, philosophical, and interdisciplinary perspectives have claimed the importance of multisensory signals and neural body representations for general theories of self-consciousness. Influential theories stated that very basic, and largely implicit and pre-reflective bodily processes crucially underlie the self (Alsmith 2012; e.g., Blanke & Metzinger 2009; Blanke 2012; Gallagher 2005; Legrand 2007). Such theories fueled experimental investigations on multisensory integration and its influence on various aspects of the self. Yet, similarly to Aristotle, who claimed that “there is no sixth sense in addition to the five enumerated—sight, hearing, smell, taste and touch”—this line of research has largely neglected the *vestibular sense of balance*. This is particularly surprising as a recent theory has claimed the importance of more global aspects of the bodily self (Blanke & Metzinger 2009), most importantly probably the sense of immersion or location in a spatiotemporal frame of reference (Windt 2010). This process, as we speculated above, should fundamentally rely on vestibular cues, plausibly among others coded by specific cells in the hippocampus. The vestibular system is activated by gravity, the constant force under which we have evolved, and also during all sorts of passive and active head and whole body movements. Moving in an environment is necessary for the development of a sense of bodily self, and the vestibular system is thus likely to contribute not only to the most basic (or minimal) aspects of the self but also to the different fine-grained implicit and explicit aspects of the experience of

our bodily self in daily life such as body perception, body ownership, agency, and self-other distinction. It is thus not surprising that the vestibular system is intrinsically, highly linked to other sensory systems such as touch, pain, interoception, and proprioception. While some of the links between the vestibular system and the bodily self are rather well-established and the underlying neurophysiological processes known from both non-human animal and human research, several of the relations presented here are still largely speculative. Yet, we believe that the specific and testable hypotheses we have given here—once they are tested and possibly confirmed by experimental studies—might enable us to better describe neural and physiological mechanisms underlying minimal phenomenal selfhood (Blanke & Metzinger 2009) as well as refine current models of the multisensory mechanisms underlying the various aspects of the bodily self.

Acknowledgements

We thank Gianluca Macaudo for his help with figures 1 and 2, as well as Dr. Jane Aspell for proofreading and her valuable comments. BL was funded by the Swiss National Science Foundation (grant #142601). CL is supported by the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant agreement number 333607 (“*BODILY-SELF, vestibular and multisensory investigations of bodily self-consciousness*”).

References

- Alsmith, A. (2012). What reason could there be to believe in pre-reflective bodily self-consciousness. In F. Paglieri (Ed.) *Consciousness in interaction: The role of the natural and social environment in shaping consciousness* (pp. 1-21). Amsterdam, NL: John Benjamins Publishing Company.
- André, J. M., Martinet, N., Paysant, J., Beis, J. M. & Le Chapelain, L. (2001). Temporary phantom limbs evoked by vestibular caloric stimulation in amputees. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology*, 14 (3), 190-196.
- Angelaki, D. E. & Cullen, K. E. (2008). Vestibular system: The many facets of a multimodal sense. *Annual Review of Neuroscience*, 31, 125-150. [10.1146/annurev.neuro.31.060407.12555](https://doi.org/10.1146/annurev.neuro.31.060407.12555)
- Angelaki, D. E., Shaikh, A. G., Green, A. M. & Dickman, J. D. (2004). Neurons compute internal models of the physical laws of motion. *Nature*, 430 (6999), 560-564. [10.1038/nature02754](https://doi.org/10.1038/nature02754)
- Armell, K. C. & Ramachandran, V. S. (2003). Projecting sensations to external objects: Evidence from skin conductance response. *Proceedings of the Royal Society B: Biological Sciences*, 270 (1523), 1499-1506. [10.1098/rspb.2003.2364](https://doi.org/10.1098/rspb.2003.2364)
- Aspell, J. E., Heydrich, L., Marillier, G., Lavanchy, T., Herbelin, B. & Blanke, O. (2013). Turning body and self inside out: Visualized heartbeats alter bodily self-consciousness and tactile perception. *Psychological Science*, 24 (12), 2445-2453. [10.1177/0956797613498395](https://doi.org/10.1177/0956797613498395)
- Aspell, J. E., Lenggenhager, B. & Blanke, O. (2009). Keeping in touch with one's self: Multisensory mechanisms of self-consciousness. *PLoS One*, 4 (8), e6488-e6488. [10.1371/journal.pone.0006488](https://doi.org/10.1371/journal.pone.0006488)
- Bahrack, L. E. & Watson, J. S. (1985). Detection of intermodal proprioceptive-visual contingency as a potential basis of self-perception in infancy. *Developmental Psychology*, 21 (6), 963-973. [10.1037/0012-1649.21.6.963](https://doi.org/10.1037/0012-1649.21.6.963)
- Balaban, C. D. (1999). Vestibular autonomic regulation (including motion sickness and the mechanism of vomiting). *Current Opinion in Neurology*, 12 (1), 29-33.
- (2004). Projections from the parabrachial nucleus to the vestibular nuclei: Potential substrates for autonomic and limbic influences on vestibular responses. *Brain Research*, 996 (1), 126-137. [10.1016/j.brainres.2003.10.026](https://doi.org/10.1016/j.brainres.2003.10.026)
- (2011). Migraine, vertigo and migrainous vertigo: Links between vestibular and pain mechanisms. *Journal of Vestibular Research*, 21 (6), 315-321. [10.3233/VES-2011-0428](https://doi.org/10.3233/VES-2011-0428)

- Barmack, N. H. (2003). Central vestibular system: Vestibular nuclei and posterior cerebellum. *Brain Research Bulletin*, 60 (5-6), 511-541. [10.1016/S0361-9230\(03\)00055-8](https://doi.org/10.1016/S0361-9230(03)00055-8)
- Barnsley, N., McAuley, J. H., Mohan, R., Dey, A., Thomas, P. & Moseley, G. L. (2011). The rubber hand illusion increases histamine reactivity in the real arm. *Current Biology*, 21 (23), R945-R946. [10.1016/j.cub.2011.10.039](https://doi.org/10.1016/j.cub.2011.10.039)
- Barra, J., Marquer, A., Joassin, R., Reymond, C., Metge, L., Chauvineau, V. & Perennou, D. (2010). Humans use internal models to construct and update a sense of verticality. *Brain*, 133, 3552-3563. [10.1093/brain/awq311](https://doi.org/10.1093/brain/awq311)
- Barry, C. & Burgess, N. (2014). Neural mechanisms of self-location. *Current Biology*, 24 (8), R330-R339. [10.1016/j.cub.2014.02.049](https://doi.org/10.1016/j.cub.2014.02.049)
- Baumgartner, U., Iannetti, G. D., Zambreanu, L., Stoeter, P., Treede, R. D. & Tracey, I. (2010). Multiple somatotopic representations of heat and mechanical pain in the operculo-insular cortex: A high-resolution fMRI study. *Journal of Neurophysiology*, 104 (5), 2863-2872. [10.1152/jn.00253.2010](https://doi.org/10.1152/jn.00253.2010)
- Bekrater-Bodmann, R., Foell, J., Diers, M., Kamping, S., Rance, M., Kirsch, P. & Flor, H. (2014). The importance of synchrony and temporal order of visual and tactile input for illusory limb ownership experiences - An fMRI study applying virtual reality. *PLoS One*, 9 (1), e87013-e87013. [10.1371/journal.pone.0087013](https://doi.org/10.1371/journal.pone.0087013)
- Bense, S., Stephan, T., Yousry, T. A., Brandt, T. & Dieterich, M. (2001). Multisensory cortical signal increases and decreases during vestibular galvanic stimulation (fMRI). *Journal of Neurophysiology*, 85 (2), 886-899.
- Berlucchi, G. & Aglioti, S. M. (2010). The body in the brain revisited. *Experimental Brain Research*, 200 (1), 25-35. [10.1007/s00221-009-1970-7](https://doi.org/10.1007/s00221-009-1970-7)
- Bermúdez, J. L. (1998). *The paradox of self-consciousness*. Cambridge, MA: MIT Press.
- Berthoz, A. (1996). How does the cerebral cortex process and utilize vestibular signals? In R. W. Baloh & G. M. Halmagyi (Eds.) *Disorders of the vestibular system* (pp. 113-125). New York, NY: Oxford University Press.
- (2000). *The brain's sense of movement*. Cambridge, MA: Harvard University Press.
- Bester, H., Menendez, L., Besson, J. M. & Bernard, J. (1995). Spino (trigemino) parabrachiohypothalamic pathway: Electrophysiological evidence for an involvement in pain processes. *Journal of Neurophysiology*, 73 (2), 568-585.
- Bisiach, E., Rusconi, M. L. & Vallar, G. (1991). Remission of somatoparaphrenic delusion through vestibular stimulation. *Neuropsychologia*, 29 (10), 1029-1031. [10.1016/0028-3932\(91\)90066-H](https://doi.org/10.1016/0028-3932(91)90066-H)
- Blanke, O. & Metzinger, T. (2009). Full-body illusions and minimal phenomenal selfhood. *Trends in Cognitive Neuroscience*, 13 (1), 7-13. [10.1016/j.tics.2008.10.003](https://doi.org/10.1016/j.tics.2008.10.003)
- Blanke, O. & Mohr, C. (2005). Out-of-body experience, heautoscopy, and autoscopic hallucination of neurological origin: Implications for neurocognitive mechanisms of corporeal awareness and self-consciousness. *Brain Research Reviews*, 50 (1), 184-199. [10.1016/j.brainresrev.2005.05.008](https://doi.org/10.1016/j.brainresrev.2005.05.008)
- Blanke, O. (2012). Multisensory brain mechanisms of bodily self-consciousness. *Nature Reviews Neuroscience*, 13 (8), 556-571. [10.1038/nrn3292](https://doi.org/10.1038/nrn3292)
- Blanke, O., Landis, T., Spinelli, L. & Seeck, M. (2004). Out-of-body experience and autoscopia of neurological origin. *Brain*, 127 (2), 243-258. [10.1093/brain/awh040](https://doi.org/10.1093/brain/awh040)
- Blanke, O., Mohr, C., Michel, C. M., Pascual-Leone, A., Brugger, P., Seeck, M., Landis, T. & Thut, G. (2005). Linking out-of-body experience and self processing to mental own-body imagery at the temporoparietal junction. *Journal of Neuroscience*, 25 (3), 550-557. [10.1523/JNEUROSCI.2612-04.2005](https://doi.org/10.1523/JNEUROSCI.2612-04.2005)
- Blanke, O., Ortigue, S., Landis, T. & Seeck, M. (2002). Stimulating illusory own-body perceptions. *Nature*, 419 (6904), 269-270. [10.1038/419269a](https://doi.org/10.1038/419269a)
- Boccarda, C. N., Sargolini, F., Thoresen, V. H., Solstad, T., Witter, M., Moser, E. I. & Moser, M. B. (2010). Grid cells in pre- and parasubiculum. *Nature Neuroscience*, 13 (8), 987-994. [10.1038/nn.2602](https://doi.org/10.1038/nn.2602)
- Bonnier, P. (1905). L'Aschématie. *Revue Neurologique*, 13, 605-609.
- Bottini, G., Karnath, H. O., Vallar, G., Sterzi, R., Frith, C. D., Frackowiak, R. S. & Paulesu, E. (2001). Cerebral representations for egocentric space: Functional-anatomical evidence from caloric vestibular stimulation and neck vibration. *Brain*, 124 (6), 1182-1196. [10.1093/brain/124.6.1182](https://doi.org/10.1093/brain/124.6.1182)
- Bottini, G., Paulesu, E., Gandola, M., Loffredo, S., Scarpa, P., Sterzi, R., Santilli, I., Defanti, C., Scialfa, G., Fazio, F. & Vallar, G. (2005). Left caloric vestibular stimulation ameliorates right hemianesthesia. *Neurology*, 65 (8), 1278-1283. [10.1212/01.wnl.0000182398.14088.e8](https://doi.org/10.1212/01.wnl.0000182398.14088.e8)
- Bottini, G., Paulesu, E., Sterzi, R., Warburton, E., Wise, R. J., Vallar, G., Frackowiak, R. & Frith, C. D. (1995). Modulation of conscious experience by peripheral sens-

- ory stimuli. *Nature*, 376 (6543), 778-781. [10.1038/376778a0](https://doi.org/10.1038/376778a0)
- Bottini, G., Sterzi, R., Paulesu, E., Vallar, G., Cappa, S. F., Erminio, F., Passingham, R., Frith, C. & Frackowiak, R. S. (1994). Identification of the central vestibular projections in man: A positron emission tomography activation study. *Experimental Brain Research*, 99 (1), 164-169. [10.1007/BF00241421](https://doi.org/10.1007/BF00241421)
- Botvinick, M. & Cohen, J. (1998). Rubber hands “feel” touch that eyes see. *Nature*, 391 (6669), 756-756. [10.1038/35784](https://doi.org/10.1038/35784)
- Brandt, T., Schautzer, F., Hamilton, D. A., Bruning, R., Markowitsch, H. J., Kalla, R., Darlington, C., Smith, P. & Strupp, M. (2005). Vestibular loss causes hippocampal atrophy and impaired spatial memory in humans. *Brain*, 128 (11), 2732-2741. [10.1093/brain/awh617](https://doi.org/10.1093/brain/awh617)
- Bremmer, F., Klam, F., Duhamel, J. R., Ben Hamed, S. & Graf, W. (2002). Visual-vestibular interactive responses in the macaque ventral intraparietal area (VIP). *European Journal of Neuroscience*, 16 (8), 1569-1586. [10.1046/j.1460-9568.2002.02206.x](https://doi.org/10.1046/j.1460-9568.2002.02206.x)
- Bremmer, F., Kubischik, M., Pekel, M., Lappe, M. & Hoffmann, K. P. (1999). Linear vestibular self-motion signals in monkey medial superior temporal area. *Annals of the New York Academy of Sciences*, 871 (1), 272-281. [10.1111/j.1749-6632.1999.tb09191.x](https://doi.org/10.1111/j.1749-6632.1999.tb09191.x)
- Bremmer, F., Schlack, A., Duhamel, J. R., Graf, W. & Fink, G. R. (2001). Space coding in primate posterior parietal cortex. *NeuroImage*, 14 (1), 46-51. [10.1006/nimg.2001.0817](https://doi.org/10.1006/nimg.2001.0817)
- Bresciani, J. P., Blouin, J., Popov, K., Bourdin, C., Sarlegna, F., Vercher, J. L. & Gauthier, G. M. (2002). Galvanic vestibular stimulation in humans produces online arm movement deviations when reaching towards memorized visual targets. *Neuroscience Letters*, 318 (1), 34-38. [10.1016/S0304-3940\(01\)02462-4](https://doi.org/10.1016/S0304-3940(01)02462-4)
- Brugger, P., Agosti, R., Regard, M., Wieser, H. G. & Landis, T. (1994). Heautoscopy, epilepsy, and suicide. *Journal of Neurology, Neurosurgery, and Psychiatry*, 57 (7), 838-839. [10.1136/jnnp.57.7.838](https://doi.org/10.1136/jnnp.57.7.838)
- Bufalari, I., Lenggenhager, B., Porciello, G., Serra Holmes, B. & Aglioti, S. M. (2014). Enfacing others but only if they are nice to you. *Frontiers in Behavioral Neuroscience*, 8 (102), 1-12. [10.3389/fnbeh.2014.00102](https://doi.org/10.3389/fnbeh.2014.00102)
- Bunning, S. & Blanke, O. (2005). The out-of body experience: Precipitating factors and neural correlates. *Progress in Brain Research*, 150, 331-350. [10.1016/S0079-6123\(05\)50024-4](https://doi.org/10.1016/S0079-6123(05)50024-4)
- Burgess, N. & O’Keefe, J. (2003). Neural representations in human spatial memory. *Trends in Cognitive Sciences*, 7 (12), 517-519. [10.1016/j.tics.2003.10.014](https://doi.org/10.1016/j.tics.2003.10.014)
- Candidi, M., Micarelli, A., Viziano, A., Aglioti, S. M., Minio-Paluello, I. & Alessandrini, M. (2013). Impaired mental rotation in benign paroxysmal positional vertigo and acute vestibular neuritis. *Frontiers in Human Neuroscience*, 7, 1-11. [10.3389/fnhum.2013.00783](https://doi.org/10.3389/fnhum.2013.00783)
- Capelari, E. D., Uribe, C. & Brasil-Neto, J. P. (2009). Feeling pain in the rubber hand: Integration of visual, proprioceptive, and painful stimuli. *Perception*, 38 (1), 92-99. [10.1068/p5892](https://doi.org/10.1068/p5892)
- Carruthers, G. (2008). Types of body representation and the sense of embodiment. *Consciousness and Cognition*, 17 (4), 1302-1316. [10.1016/j.concog.2008.02.001](https://doi.org/10.1016/j.concog.2008.02.001)
- Chen, A., DeAngelis, G. C. & Angelaki, D. E. (2010). Macaque parieto-insular vestibular cortex: Responses to self-motion and optic flow. *Journal of Neuroscience*, 30 (8), 3022-3042. [10.1523/JNEUROSCI.4029-09.2010](https://doi.org/10.1523/JNEUROSCI.4029-09.2010)
- (2011). A comparison of vestibular spatiotemporal tuning in macaque parietoinsular vestibular cortex, ventral intraparietal area, and medial superior temporal area. *Journal of Neuroscience*, 31 (8), 3082-3094. [10.1523/JNEUROSCI.4476-10.2011](https://doi.org/10.1523/JNEUROSCI.4476-10.2011)
- Cole, J. (2004). *Still lives*. Cambridge, MA: MIT Press.
- Coq, J. Q., Qi, H., Collins, C. E. & Kaas, J. H. (2004). Anatomical and functional organization of somatosensory areas of the lateral fissure of the new world titi monkey (*Callicebus moloch*). *The Journal of Comparative Neurology*, 476 (4), 363-387. [10.1002/cne.20237](https://doi.org/10.1002/cne.20237)
- Corradi-Dell’acqua, C., Ueno, K., Ogawa, A., Cheng, K., Rumiati, R. I. & Iriki, A. (2008). Effects of shifting perspective of the self: An fMRI study. *NeuroImage*, 40 (4), 1902-1911. [10.1016/j.neuroimage.2007.12.062](https://doi.org/10.1016/j.neuroimage.2007.12.062)
- Courtial, E. & Wilson, D. A. (2014). Thalamic olfaction: Characterizing odor processing in the mediodorsal thalamus of the rat. *Journal of Neurophysiology*, 111 (6), 1274-1285. [10.1152/jn.00741.2013](https://doi.org/10.1152/jn.00741.2013)
- Craig, A. D. (2009). How do you feel-now? The anterior insula and human awareness. *Nature Reviews Neuroscience*, 10, 59-70. [10.1038/nrn2555](https://doi.org/10.1038/nrn2555)
- Cullen, K. E. (2011). The neural encoding of self-motion. *Current Opinion in Neurobiology*, 21 (4), 587-595. [10.1016/j.conb.2011.05.022](https://doi.org/10.1016/j.conb.2011.05.022)
- (2012). The vestibular system: Multimodal integration and encoding of self-motion for motor control.

- Trends in Neurosciences*, 35 (3), 185-196.
[10.1016/j.tins.2011.12.001](https://doi.org/10.1016/j.tins.2011.12.001)
- Cullen, K. E., Roy, J. E. & Sylvestre, P. A. (2003). Signal processing in vestibular nuclei: Dissociating sensory, motor, and cognitive influence. In L. Harris & M. Jenkin (Eds.) *Levels of perception* (pp. 285-309). Berlin, GER: Springer.
- David, N., Bewernick, B. H., Cohen, M. X., Newen, A., Lux, S., Fink, G. R. & Vogeley, K. (2006). Neural representations of self versus other: Visual-spatial perspective taking and agency in a virtual ball-tossing game. *Journal of Cognitive Neuroscience*, 18 (6), 898-910. [10.1162/jocn.2006.18.6.898](https://doi.org/10.1162/jocn.2006.18.6.898)
- David, N., Fiori, F. & Aglioti, S. M. (2014). Susceptibility to the rubber hand illusion does not tell the whole body-awareness story. *Cognitive, Affective & Behavioral Neuroscience*, 14 (1), 297-306.
[10.3758/s13415-013-0190-6](https://doi.org/10.3758/s13415-013-0190-6)
- Day, B. L. & Fitzpatrick, R. C. (2005). The vestibular system. *Current Biology*, 15 (15), R583-R586.
[10.1016/j.cub.2005.07.053](https://doi.org/10.1016/j.cub.2005.07.053)
- De Vignemont, F. (2010). Body schema and body image-pros and cons. *Neuropsychologia*, 48 (3), 669-680.
[10.1016/j.neuropsychologia.2009.09.022](https://doi.org/10.1016/j.neuropsychologia.2009.09.022)
- Deroualle, D. & Lopez, C. (2014). Toward a vestibular contribution to social cognition. *Frontiers in Integrative Neuroscience*, 8, 1-4. [10.3389/fnint.2014.00016](https://doi.org/10.3389/fnint.2014.00016)
- Dichgans, J. & Brandt, T. (1978). Visual-vestibular interaction: Effects on self-motion perception and postural control. In H. Autrum, R. Jung, W. R. Loewenstein, D. M. MacKay & H. L. Teuber (Eds.) *Handbook of sensory physiology* (pp. 755-804). Berlin, GER: Springer.
- Dieterich, M., Bense, S., Lutz, S., Drzezga, A., Stephan, T., Bartenstein, P. & Brandt, T. (2003). Dominance for vestibular cortical function in the non-dominant hemisphere. *Cerebral Cortex*, 13 (9), 994-1007.
[10.1093/cercor/13.9.994](https://doi.org/10.1093/cercor/13.9.994)
- Dilda, V., MacDougall, H. G., Curthoys, I. S. & Moore, S. T. (2012). Effects of galvanic vestibular stimulation on cognitive function. *Experimental Brain Research*, 216 (2), 275-285. [10.1007/s00221-011-2929-z](https://doi.org/10.1007/s00221-011-2929-z)
- Dobricki, M. & de la Rosa, S. (2013). The structure of conscious bodily self-perception during full-body illusions. *PLoS One*, 8 (12), e83840-e83840.
[10.1371/journal.pone.0083840](https://doi.org/10.1371/journal.pone.0083840)
- Doeller, C. F., Barry, C. & Burgess, N. (2010). Evidence for grid cells in a human memory network. *Nature*, 463 (7281), 657-661. [10.1038/nature08704](https://doi.org/10.1038/nature08704)
- Ebata, S., Sugiuchi, Y., Izawa, Y., Shinomiya, K. & Shinoda, Y. (2004). Vestibular projection to the periaudate cortex in the monkey. *Neuroscience Research*, 49 (1), 55-68. [10.1016/j.neures.2004.01.012](https://doi.org/10.1016/j.neures.2004.01.012)
- Ehrsson, H. H. (2007). The experimental induction of out-of-body experiences. *Science*, 317 (5841), 1048-1048. [10.1126/science.1142175](https://doi.org/10.1126/science.1142175)
- Ehrsson, H. H., Spence, C. & Passingham, R. E. (2004). That's my hand! Activity in premotor cortex reflects feeling of ownership of a limb. *Science*, 305 (5685), 875-877. [10.1126/science.1097011](https://doi.org/10.1126/science.1097011)
- Ehrsson, H. H., Wiech, K., Weiskopf, N., Dolan, R. J. & Passingham, R. E. (2007). Threatening a rubber hand that you feel is yours elicits a cortical anxiety response. *Proceedings of the National Academy of Sciences of the United States of America*, 104 (23), 9828-9833.
[10.1073/pnas.0610011104](https://doi.org/10.1073/pnas.0610011104)
- Eickhoff, S. B., Weiss, P. H., Amunts, K., Fink, G. R. & Zilles, K. (2006). Identifying human parieto-insular vestibular cortex using fMRI and cytoarchitectonic mapping. *Human Brain Mapping*, 27 (7), 611-621.
[10.1002/hbm.20205](https://doi.org/10.1002/hbm.20205)
- Ekstrom, A. D., Kahana, M. J., Caplan, J. B., Fields, T. A., Isham, E. A., Newman, E. L. & Fried, I. (2003). Cellular networks underlying human spatial navigation. *Nature*, 425 (6954), 184-188.
[10.1038/nature01964](https://doi.org/10.1038/nature01964)
- Falconer, C. J. & Mast, F. W. (2012). Balancing the mind: Vestibular induced facilitation of egocentric mental transformations. *Experimental Psychology*, 59 (6), 332-339. [10.1027/1618-3169/a000161](https://doi.org/10.1027/1618-3169/a000161)
- Farrer, C., Franck, N., Georgieff, N., Frith, C. D., Decety, J. & Jeannerod, M. (2003a). Modulating the experience of agency: A positron emission tomography study. *NeuroImage*, 18 (2), 324-333.
[10.1016/S1053-8119\(02\)00041-1](https://doi.org/10.1016/S1053-8119(02)00041-1)
- Farrer, C., Franck, N., Paillard, J. & Jeannerod, M. (2003b). The role of proprioception in action recognition. *Consciousness and Cognition*, 12 (4), 609-619.
[10.1016/S1053-8100\(03\)00047-3](https://doi.org/10.1016/S1053-8100(03)00047-3)
- Farrer, C., Frey, S. H., Van Horn, J. D., Tunik, E., Turk, D., Inati, S. & Grafton, S. T. (2008). The angular gyrus computes action awareness representations. *Cerebral Cortex*, 18, 254-261.
[10.1016/S1053-8100\(03\)00047-3](https://doi.org/10.1016/S1053-8100(03)00047-3)
- Fasold, O., von Brevern, M., Kuhberg, M., Ploner, C. J., Villringer, A., Lempert, T. & Wenzel, R. (2002). Human vestibular cortex as identified with caloric stimulation in functional magnetic resonance imaging.

- NeuroImage*, 17 (3), 1384-1393.
[10.1006/nimg.2002.1241](https://doi.org/10.1006/nimg.2002.1241)
- Feil, K. & Herbert, H. (1995). Topographic organization of spinal and trigeminal somatosensory pathways to the rat parabrachial and Kölliker-Fuse nuclei. *The Journal of Comparative Neurology*, 353 (4), 506-528.
[10.1002/cne.903530404](https://doi.org/10.1002/cne.903530404)
- Fernández, C. & Goldberg, J. (1976). Physiology of peripheral neurons innervating otolith organs of the squirrel monkey. I. Response to static tilts and to long-duration centrifugal force. *Journal of Neurophysiology*, 39 (5), 970-984.
- Ferrè, E. R., Bottini, G. & Haggard, P. (2011). Vestibular modulation of somatosensory perception. *The European Journal of Neuroscience*, 34 (8), 1337-1344.
[10.1111/j.1460-9568.2011.07859.x](https://doi.org/10.1111/j.1460-9568.2011.07859.x)
- (2012). Vestibular inputs modulate somatosensory cortical processing. *Brain Structure & Function*, 217 (4), 859-864. [10.1007/s00429-012-0404-7](https://doi.org/10.1007/s00429-012-0404-7)
- Ferrè, E. R., Vagnoni, E. & Haggard, P. (2013). Vestibular contributions to bodily awareness. *Neuropsychologia*, 51 (8), 1445-1452.
[10.1016/j.neuropsychologia.2013.04.006](https://doi.org/10.1016/j.neuropsychologia.2013.04.006)
- Filippetti, M. L., Johnson, M. H., Lloyd-Fox, S., Dragovic, D. & Farroni, T. (2013). Body perception in newborns. *Current Biology*, 23 (23), 2413-2416.
[10.1016/j.cub.2013.10.017](https://doi.org/10.1016/j.cub.2013.10.017)
- Fitzpatrick, R. C. & Day, B. L. (2004). Probing the human vestibular system with galvanic stimulation. *Journal of Applied Physiology*, 96 (6), 2301-2316.
[10.1152/jappphysiol.00008.2004](https://doi.org/10.1152/jappphysiol.00008.2004)
- Fitzpatrick, R. C., Marsden, J., Lord, S. R. & Day, B. L. (2002). Galvanic vestibular stimulation evokes sensations of body rotation. *Neuroreport*, 13 (18), 2379-2383.
[10.1097/01.wnr.0000048002.96487.de](https://doi.org/10.1097/01.wnr.0000048002.96487.de)
- Fourneret, P. & Jeannerod, M. (1998). Limited conscious monitoring of motor performance in normal subjects. *Neuropsychologia*, 36 (11), 1133-1140.
[10.1016/S0028-3932\(98\)00006-2](https://doi.org/10.1016/S0028-3932(98)00006-2)
- Fourneret, P., Franck, N., Slachevsky, A. & Jeannerod, M. (2001). Self-monitoring in schizophrenia revisited. *NeuroReport*, 12 (6), 1203-1208.
[10.1097/00001756-200105080-00030](https://doi.org/10.1097/00001756-200105080-00030)
- Friston, K. (2012). Predictive coding, precision and synchrony. *Cognitive Neuroscience*, 3 (3-4), 238-239.
[10.1080/17588928.2012.691277](https://doi.org/10.1080/17588928.2012.691277)
- Fukushima, K. (1997). Cortico-vestibular interactions: Anatomy, electrophysiology, and functional considerations. *Experimental Brain Research*, 117 (1), 1-16.
[10.1007/PL00005786](https://doi.org/10.1007/PL00005786)
- Fukushima, J., Akao, T., Kurkin, S., Kaneko, C. R. & Fukushima, K. (2006). The vestibular-related frontal cortex and its role in smooth-pursuit eye movements and vestibular-pursuit interactions. *Journal of Vestibular Research*, 16 (1-2), 1-22.
- Furlanetto, T., Bertone, C. & Becchio, C. (2013). The bi-located mind: New perspectives on self-localization and self-identification. *Frontiers in Human Neuroscience*, 7, 1-6. [10.3389/fnhum.2013.00071](https://doi.org/10.3389/fnhum.2013.00071)
- Furman, D. J., Waugh, C. E., Bhattacharjee, K., Thompson, R. J. & Gotlib, I. H. (2013). Interoceptive awareness, positive affect, and decision making in Major Depressive Disorder. *Journal of Affective Disorders*, 151 (2), 780-785. [10.1016/j.jad.2013.06.044](https://doi.org/10.1016/j.jad.2013.06.044)
- Furuya, Y., Matsumoto, J., Hori, E., Boas, C. V., Tran, A. H., Shimada, Y. & Nishijo, H. (2014). Place-related neuronal activity in the monkey parahippocampal gyrus and hippocampal formation during virtual navigation. *Hippocampus*, 24 (1), 113-130.
[10.1002/hipo.22209](https://doi.org/10.1002/hipo.22209)
- Fyhn, M., Hafting, T., Witter, M. P., Moser, E. & Moser, M. B. (2008). Grid cells in mice. *Hippocampus*, 18 (12), 1230-1238. [10.1002/hipo.20472](https://doi.org/10.1002/hipo.20472)
- Gallagher, S. (2000). Philosophical conceptions of the self: Implications for cognitive science. *Trends in Cognitive Neuroscience*, 4 (1), 14-21.
[10.1016/S1364-6613\(99\)01417-5](https://doi.org/10.1016/S1364-6613(99)01417-5)
- (2005). *How the body shapes the mind*. New York, NY: Oxford University Press.
- Gallese, V., Fadiga, L., Fogassi, L. & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, 119 (2), 593-609. [10.1093/brain/119.2.593](https://doi.org/10.1093/brain/119.2.593)
- Ganesh, S., van Schie, H. T., de Lange, F. P., Thompson, E. & Wigboldus, D. H. J. (2012). How the human brain goes virtual: Distinct cortical regions of the person-processing network are involved in self-identification with virtual agents. *Cerebral Cortex*, 22 (7), 1577-1585. [10.1093/cercor/bhr227](https://doi.org/10.1093/cercor/bhr227)
- Gentsch, A. & Schütz-Bosbach, S. (2011). I did it: Unconscious expectation of sensory consequences modulates the experience of self-agency and its functional signature. *Journal of Cognitive Neuroscience*, 23 (12), 3817-3828. [10.1162/jocn_a_00012](https://doi.org/10.1162/jocn_a_00012)
- Gilbert, J. W., Vogt, M., Windsor, R. E., Mick, G. E., Richardson, G. B., Storey, B. B. & Maddox, M. L. (2014). Vestibular dysfunction in patients with chronic pain or underlying neurologic disorders. *The Journal of the American Osteopathic Association*, 114 (3), 172-178. [10.7556/jaoa.2014.034](https://doi.org/10.7556/jaoa.2014.034)

- Glasauer, S., Amorim, M. A., Vitte, E. & Berthoz, A. (1994). Goal-directed linear locomotion in normal and labyrinthine-defective subjects. *Experimental Brain Research*, 98 (2), 323-335. [10.1007/BF00228420](https://doi.org/10.1007/BF00228420)
- Green, C. (1968). *Out-of-body experiences*. Oxford, UK: Institute of Psychophysical Research.
- Grüsser, O. J., Guldin, W. O., Mirring, S. & Salah-Eldin, A. (1994). Comparative physiological and anatomical studies of the primate vestibular cortex. In B. Al-bowitz, K. Albus, U. Kuhnt, H. C. Nothdurf & P. Wahle (Eds.) *Structural and functional organization of the neocortex* (pp. 358-371). Berlin, GER: Springer.
- Grüsser, O. J., Pause, M. & Schreier, U. (1990a). Localization and responses of neurones in the parieto-insular vestibular cortex of awake monkeys (*Macaca fascicularis*). *Journal of Physiology*, 430, 537-557.
- (1990b). Vestibular neurones in the parieto-insular cortex of monkeys (*Macaca fascicularis*): Visual and neck receptor responses. *Journal of Physiology*, 430, 559-583.
- Gu, Y., DeAngelis, G. C. & Angelaki, D. E. (2007). A functional link between area MSTd and heading perception based on vestibular signals. *Nature Neuroscience*, 10 (8), 1038-1047. [10.1038/nn1935](https://doi.org/10.1038/nn1935)
- Guillaud, E., Simoneau, M. & Blouin, J. (2011). Prediction of the body rotation-induced torques on the arm during reaching movements: Evidence from a proprioceptively deafferented subject. *Neuropsychologia*, 49 (7), 2055-2059. [10.1016/j.neuropsychologia.2011.03.035](https://doi.org/10.1016/j.neuropsychologia.2011.03.035)
- Guldin, W. O. & Grüsser, O. J. (1998). Is there a vestibular cortex? *Trends in Neuroscience*, 21 (6), 254-359. [10.1016/S0166-2236\(97\)01211-3](https://doi.org/10.1016/S0166-2236(97)01211-3)
- Guldin, W. O., Akbarian, S. & Grüsser, O. J. (1992). Cortico-cortical connections and cytoarchitectonics of the primate vestibular cortex: A study in squirrel monkeys (*Saimiri sciureus*). *Journal of Comparative Neurology*, 326 (3), 375-401. [10.1002/cne.903260306](https://doi.org/10.1002/cne.903260306)
- Gutiérrez-Martínez, O., Gutiérrez-Maldonado, J. & Loreto-Quijada, D. (2011). Control over the virtual environment influences the presence and efficacy of a virtual reality intervention on pain. *Studies in Health Technology and Informatics*, 167, 111-115. [10.3233/978-1-60750-766-6-111](https://doi.org/10.3233/978-1-60750-766-6-111)
- Hänsel, A., Lenggenhager, B., von Känel, R., Curatolo, M. & Blanke, O. (2011). Seeing and identifying with a virtual body decreases pain perception. *European Journal of Pain*, 15 (8), 874-879. [10.1016/j.ejpain.2011.03.013](https://doi.org/10.1016/j.ejpain.2011.03.013)
- Herbert, H., Moga, M. M. & Saper, C. B. (1990). Connections of the parabrachial nucleus with the nucleus of the solitary tract and the medullary reticular formation in the rat. *The Journal of Comparative Neurology*, 293 (4), 540-580. [10.1002/cne.902930404](https://doi.org/10.1002/cne.902930404)
- Heydrich, L. & Blanke, O. (2013). Distinct illusory own-body perceptions caused by damage to posterior insula and extrastriate cortex. *Brain*, 136 (3), 790-803. [10.1093/brain/aws364](https://doi.org/10.1093/brain/aws364)
- Hoffman, H. G., Sharar, S. R., Coda, B., Everett, J. J., Ciol, M., Richards, T. & Patterson, D. R. (2004). Manipulating presence influences the magnitude of virtual reality analgesia. *Pain*, 111 (1-2), 162-168. [10.1016/j.pain.2004.06.013](https://doi.org/10.1016/j.pain.2004.06.013)
- Hüfner, K., Hamilton, D. A., Kalla, R., Stephan, T., Glasauer, S., Ma, J. & Brandt, T. (2007). Spatial memory and hippocampal volume in humans with unilateral vestibular deafferentation. *Hippocampus*, 17 (6), 471-485. [10.1002/hipo.20283](https://doi.org/10.1002/hipo.20283)
- Indovina, I., Maffei, V., Bosco, G., Zago, M., Macaluso, E. & Lacquaniti, F. (2005). Representation of visual gravitational motion in the human vestibular cortex. *Science*, 308 (5720), 416-419. [10.1126/science.1107961](https://doi.org/10.1126/science.1107961)
- Ionta, S., Heydrich, L., Lenggenhager, B., Mouthon, M., Fornari, E., Chapuis, D. & Blanke, O. (2011). Multisensory mechanisms in temporo-parietal cortex support self-location and first-person perspective. *Neuron*, 70 (2), 363-374. [10.1016/j.neuron.2011.03.009](https://doi.org/10.1016/j.neuron.2011.03.009)
- Ionta, S., Perruchoud, D., Draganski, B. & Blanke, O. (2012). Body context and posture affect mental imagery of hands. *PLoS One*, 7 (3), e34382-e34382. [10.1371/journal.pone.0034382](https://doi.org/10.1371/journal.pone.0034382)
- Jacobs, J., Weidemann, C. T., Miller, J. F., Solway, A., Burke, J. F., Wei, X. X. & Kahana, M. J. (2013). Direct recordings of grid-like neuronal activity in human spatial navigation. *Nature Neuroscience*, 16 (9), 1188-1190. [10.1038/nn.3466](https://doi.org/10.1038/nn.3466)
- Jasmin, L., Burke, A. R., Card, J. P. & Basbaum, A. I. (1997). Transneuronal labeling of a nociceptive pathway, the spino-(trigemino-)parabrachio-amygdaloid, in the rat. *Journal of Neuroscience*, 17 (10), 3751-3765.
- Jauregui-Renaud, K., Villanueva, P. L. & del Castillo, M. S. (2005). Influence of acute unilateral vestibular lesions on the respiratory rhythm after active change of posture in human subjects. *Journal of Vestibular Research*, 15 (1), 41-48.
- Jeannerod, M. (2003). The mechanism of self-recognition in humans. *Behavioural Brain Research*, 142 (1-2), 1-15. [10.1016/S0166-4328\(02\)00384-4](https://doi.org/10.1016/S0166-4328(02)00384-4)

- (2006). *Motor cognition: What actions tell to the self*. Oxford, UK: Oxford University Press.
- Kannape, O. A., Schwabe, L., Tadi, T. & Blanke, O. (2010). The limits of agency in walking humans. *Neuropsychologia*, 48 (6), 1628-1636. [10.1016/j.neuropsychologia.2010.02.005](https://doi.org/10.1016/j.neuropsychologia.2010.02.005)
- Kavounoudias, A., Roll, R. & Roll, J. P. (1998). The plantar sole is a “dynamometric map” for human balance control. *NeuroReport*, 9 (14), 3247-3252.
- Kerkhoff, G., Hildebrandt, H., Reinhart, S., Kardinal, M., Dimova, V. & Utz, K. S. (2011). A long-lasting improvement of tactile extinction after galvanic vestibular stimulation: Two Sham-stimulation controlled case studies. *Neuropsychologia*, 49 (2), 186-195. [10.1016/j.neuropsychologia.2010.11.014](https://doi.org/10.1016/j.neuropsychologia.2010.11.014)
- Kessler, K. & Thomson, L. A. (2010). The embodied nature of spatial perspective taking: Embodied transformation versus sensorimotor interference. *Cognition*, 114 (1), 72-88. [10.1016/j.cognition.2009.08.015](https://doi.org/10.1016/j.cognition.2009.08.015)
- Keysers, C., Wicker, B., Gazzola, V., Anton, J. L., Fogassi, L. & Gallese, V. (2004). A touching sight: SII/PV activation during the observation and experience of touch. *Neuron*, 42 (2), 335-346. [10.1016/S0896-6273\(04\)00156-4](https://doi.org/10.1016/S0896-6273(04)00156-4)
- Killian, N. J., Jutras, M. J. & Buffalo, E. A. (2012). A map of visual space in the primate entorhinal cortex. *Nature*, 491 (7426), 761-764. [10.1038/nature11587](https://doi.org/10.1038/nature11587)
- Kim, Y. R., Son, J. W., Lee, S. I., Shin, C. J., Kim, S. K., Ju, G. & Ha, T. H. (2012). Abnormal brain activation of adolescent internet addict in a ball-throwing animation task: Possible neural correlates of disembodiment revealed by fMRI. *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, 39 (1), 88-95. [10.1016/j.pnpbp.2012.05.013](https://doi.org/10.1016/j.pnpbp.2012.05.013)
- Klam, F. & Graf, W. (2003). Vestibular signals of posterior parietal cortex neurons during active and passive head movements in macaque monkeys. *Annals of the New York Academy of Sciences*, 1004, 271-282. [10.1196/annals.1303.024](https://doi.org/10.1196/annals.1303.024)
- (2006). Discrimination between active and passive head movements by macaque ventral and medial intraparietal cortex neurons. *Journal of Physiology*, 574 (2), 367-386. [10.1113/jphysiol.2005.103697](https://doi.org/10.1113/jphysiol.2005.103697)
- Kotchabhakdi, N., Rinvik, E., Walberg, F. & Yingchareon, K. (1980). The vestibulo-thalamic projections in the cat studied by retrograde axonal transport of horseradish peroxidase. *Experimental Brain Research*, 40 (4), 405-418. [10.1007/BF00236149](https://doi.org/10.1007/BF00236149)
- Kurth, F., Zilles, K., Fox, P. T., Laird, A. R. & Eickhoff, S. B. (2010). A link between the systems: Functional differentiation and integration within the human insula revealed by meta-analysis. *Brain Structure and Function*, 214 (5-6), 519-534. [10.1007/s00429-010-0255-z](https://doi.org/10.1007/s00429-010-0255-z)
- Lackner, J. R. & DiZio, P. (2005). Vestibular, proprioceptive, and haptic contributions to spatial orientation. *Annual Reviews of Psychology*, 56, 115-147. [10.1146/annurev.psych.55.090902.142023](https://doi.org/10.1146/annurev.psych.55.090902.142023)
- Lackner, J. R. (1992). Spatial orientation in weightless environments. *Perception*, 21 (6), 803-812. [10.1068/p210803](https://doi.org/10.1068/p210803)
- Lacquaniti, F., Bosco, G., Indovina, I., La Scaleia, B., Maffei, V., Moscatelli, A. & Zago, M. (2013). Visual gravitational motion and the vestibular system in humans. *Frontiers in Integrative Neuroscience*, 7, 1-12. [10.3389/fnint.2013.00101](https://doi.org/10.3389/fnint.2013.00101)
- Lai, H., Tsumori, T., Shiroyama, T., Yokota, S., Nakano, K. & Yasui, Y. (2000). Morphological evidence for a vestibulo-thalamo-striatal pathway via the parafascicular nucleus in the rat. *Brain Research*, 872 (1-2), 208-214. [10.1016/S0006-8993\(00\)02457-4](https://doi.org/10.1016/S0006-8993(00)02457-4)
- Lambrey, S., Doeller, C., Berthoz, A. & Burgess, N. (2012). Imagining being somewhere else: Neural basis of changing perspective in space. *Cerebral Cortex*, 22 (1), 166-174. [10.1093/cercor/bhr101](https://doi.org/10.1093/cercor/bhr101)
- Lamm, C., Decety, J. & Singer, T. (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *NeuroImage*, 54 (3), 2492-2502. [10.1016/j.neuroimage.2010.10.014](https://doi.org/10.1016/j.neuroimage.2010.10.014)
- Legrand, D. (2007). Pre-reflective self-consciousness: On being bodily in the world. *Janus Head*, 9 (2), 493-519.
- Lenggenhager, B., Azevedo, R. T., Mancini, A. & Aglioti, S. M. (2013). Listening to your heart and feeling yourself: Effects of exposure to interoceptive signals during the ultimatum game. *Experimental Brain Research*, 230 (2), 233-241. [10.1007/s00221-013-3647-5](https://doi.org/10.1007/s00221-013-3647-5)
- Lenggenhager, B., Halje, P. & Blanke, O. (2011). Alpha band oscillations correlate with illusory self-location induced by virtual reality. *European Journal of Neuroscience*, 33 (10), 1935-1943. [10.1111/j.1460-9568.2011.07647.x](https://doi.org/10.1111/j.1460-9568.2011.07647.x)
- Lenggenhager, B., Lopez, C. & Blanke, O. (2008). Influence of galvanic vestibular stimulation on egocentric and object-based mental transformations. *Experimental Brain Research*, 184 (2), 211-221. [10.1007/s00221-007-1095-9](https://doi.org/10.1007/s00221-007-1095-9)

- Lenggenhager, B., Mouthon, M. & Blanke, O. (2009). Spatial aspects of bodily self-consciousness. *Consciousness and Cognition*, 18 (1), 110-117. [10.1016/j.concog.2008.11.003](https://doi.org/10.1016/j.concog.2008.11.003)
- Lenggenhager, B., Pazzaglia, M., Scivoletto, G., Molinari, M. & Aglioti, S. . (2012). The sense of the body in individuals with spinal cord injury. *PLoS ONE*, 7 (11), e50757-e50757. [10.1371/journal.pone.0050757](https://doi.org/10.1371/journal.pone.0050757)
- Lenggenhager, B., Tadi, T., Metzinger, T. & Blanke, O. (2007). Video ergo sum: Manipulating bodily self-consciousness. *Science*, 317 (5841), 1096-1099. [10.1126/science.1143439](https://doi.org/10.1126/science.1143439)
- Limanowski, J., Lutti, A. & Blankenburg, F. (2014). The extrastriate body area is involved in illusory limb ownership. *NeuroImage*, 86, 514-524. [10.1016/j.neuroimage.2013.10.035](https://doi.org/10.1016/j.neuroimage.2013.10.035)
- Liu, S., Dickman, J. D. & Angelaki, D. E. (2011). Response dynamics and tilt versus translation discrimination in parietoinsular vestibular cortex. *Cerebral Cortex*, 21 (3), 563-573. [10.1093/cercor/bhq123](https://doi.org/10.1093/cercor/bhq123)
- Lobel, E., Kleine, J. F., Le Bihan, D., Leroy-Willig, A. & Berthoz, A. (1998). Functional MRI of galvanic vestibular stimulation. *Journal of Neurophysiology*, 80 (5), 2699-2709.
- Longo, M. R. & Haggard, P. (2010). An implicit body representation underlying human position sense. *Proceedings of the National Academy of Sciences of the United States of America*, 107 (26), 11727-11732. [10.1073/pnas.1003483107](https://doi.org/10.1073/pnas.1003483107)
- Longo, M. R., Schuur, F., Kammers, M. P., Tsakiris, M. & Haggard, P. (2008). What is embodiment? A psychometric approach. *Cognition*, 107 (3), 978-998. [10.1016/j.cognition.2007.12.004](https://doi.org/10.1016/j.cognition.2007.12.004)
- Lopez, C. & Blanke, O. (2011). The thalamocortical vestibular system in animals and humans. *Brain Research Reviews*, 67 (1-2), 119-146. [10.1016/j.brainresrev.2010.12.002](https://doi.org/10.1016/j.brainresrev.2010.12.002)
- Lopez, C. (2013). A neuroscientific account of how vestibular disorders impair bodily self-consciousness. *Frontiers in Integrative Neuroscience*, 7, 1-8. [10.3389/fnint.2013.00091](https://doi.org/10.3389/fnint.2013.00091)
- Lopez, C., Blanke, O. & Mast, F. W. (2012). The vestibular cortex in the human brain revealed by coordinate-based activation likelihood estimation meta-analysis. *Neuroscience*, 212, 159-179. [10.1016/j.neuroscience.2012.03.028](https://doi.org/10.1016/j.neuroscience.2012.03.028)
- Lopez, C., Falconer, C. J. & Mast, F. W. (2013). Being moved by the self and others: Influence of empathy on self-motion perception. *PLoS One*, 8 (1), e48293-e48293. [10.1371/journal.pone.0048293](https://doi.org/10.1371/journal.pone.0048293)
- Lopez, C., Halje, P. & Blanke, O. (2008). Body ownership and embodiment: Vestibular and multisensory mechanisms. *Clinical Neurophysiology*, 38 (3), 149-161. [10.1016/j.neucli.2007.12.006](https://doi.org/10.1016/j.neucli.2007.12.006)
- Lopez, C., Lenggenhager, B. & Blanke, O. (2010). How vestibular stimulation interacts with illusory hand ownership. *Consciousness and Cognition*, 19 (1), 33-47. [10.1016/j.concog.2009.12.003](https://doi.org/10.1016/j.concog.2009.12.003)
- Lopez, C., Schreyer, H. M., Preuss, N. & Mast, F. W. (2012). Vestibular stimulation modifies the body schema. *Neuropsychologia*, 50 (8), 1830-1837. [10.1016/j.neuropsychologia.2012.04.008](https://doi.org/10.1016/j.neuropsychologia.2012.04.008)
- Ludvig, N., Tang, H. M., Gohil, B. C. & Botero, J. M. (2004). Detecting location-specific neuronal firing rate increases in the hippocampus of freely-moving monkeys. *Brain Research*, 1014 (1-2), 97-109. [10.1016/j.brainres.2004.03.071](https://doi.org/10.1016/j.brainres.2004.03.071)
- Macpherson, F. (2011). *The senses: Classic and contemporary philosophical perspectives*. Oxford, UK: Oxford University Press.
- Maister, L. & Tsakiris, M. (2014). My face, my heart: Cultural differences in integrated bodily self-awareness. *Cognitive Neuroscience*, 5 (1), 10-16. [10.1080/17588928.2013.808613](https://doi.org/10.1080/17588928.2013.808613)
- Makin, T. R., Holmes, N. P. & Ehrsson, H. H. (2008). On the other hand: Dummy hands and peripersonal space. *Behavioural Brain Research*, 191 (1), 1-10. [10.1016/j.bbr.2008.02.041](https://doi.org/10.1016/j.bbr.2008.02.041)
- Marlinski, V. & McCrea, R. A. (2008a). Activity of ventroposterior thalamus neurons during rotation and translation in the horizontal plane in the alert squirrel monkey. *Journal of Neurophysiology*, 99 (5), 2533-2545. [10.1152/jn.00761.2007](https://doi.org/10.1152/jn.00761.2007)
- (2008b). Coding of self-motion signals in ventroposterior thalamus neurons in the alert squirrel monkey. *Experimental Brain Research*, 189 (4), 463-472. [10.1007/s00221-008-1442-5](https://doi.org/10.1007/s00221-008-1442-5)
- Matsumura, N., Nishijo, H., Tamura, R., Eifuku, S., Endo, S. & Ono, T. (1999). Spatial- and task-dependent neuronal responses during real and virtual translocation in the monkey hippocampal formation. *Journal of Neuroscience*, 19 (6), 2381-2393.
- Mazzola, L., Faillenot, I., Barral, F. G., Mauguière, F. & Peyron, R. (2012). Spatial segregation of somato-sensory and pain activations in the human operculo-insular cortex. *NeuroImage*, 60 (1), 409-418. [10.1016/j.neuroimage.2011.12.072](https://doi.org/10.1016/j.neuroimage.2011.12.072)
- Mazzola, L., Isnard, J., Peyron, R., Guénot, M. & Mauguière, F. (2009). Somatotopic organization of pain

- responses to direct electrical stimulation of the human insular cortex. *Pain*, 146 (12), 99-104.
[10.1016/j.pain.2009.07.014](https://doi.org/10.1016/j.pain.2009.07.014)
- McCandless, C. H. & Balaban, C. D. (2010). Parabrachial nucleus neuronal responses to off-vertical axis rotation in macaques. *Experimental Brain Research*, 202 (2), 271-290. [10.1007/s00221-009-2130-9](https://doi.org/10.1007/s00221-009-2130-9)
- McGeoch, P. D., Williams, L. E., Lee, R. R. & Ramachandran, V. S. (2008). Behavioural evidence for vestibular stimulation as a treatment for central post-stroke pain. *Journal of Neurology, Neurosurgery and Psychiatry*, 79 (11), 1298-1301.
[10.1136/jnnp.2008.146738](https://doi.org/10.1136/jnnp.2008.146738)
- McHugh, T. J., Blum, K. I., Tsien, J. Z., Tonegawa, S. & Wilson, M. A. (1996). Impaired hippocampal representation of space in CA1-specific NMDAR1 knockout mice. *Cell*, 87 (7), 1339-1349.
[10.1016/S0092-8674\(00\)81828-0](https://doi.org/10.1016/S0092-8674(00)81828-0)
- McIntyre, J., Zago, M., Berthoz, A. & Lacquaniti, F. (2001). Does the brain model Newton's laws? *Nature Neuroscience*, 4 (7), 693-694. [10.1038/89477](https://doi.org/10.1038/89477)
- Meng, H., Bai, R. S., Sato, H., Imagawa, M., Sasaki, M. & Uchino, Y. (2001). Otolith-activated vestibulothalamic neurons in cats. *Experimental Brain Research*, 141 (4), 415-424. [10.1007/s00221-001-0902-y](https://doi.org/10.1007/s00221-001-0902-y)
- Meng, H., May, P. J., Dickman, J. D. & Angelaki, D. E. (2007). Vestibular signals in primate thalamus: Properties and origins. *Journal of Neuroscience*, 27 (50), 13590-13602. [10.1523/JNEUROSCI.3931-07.2007](https://doi.org/10.1523/JNEUROSCI.3931-07.2007)
- Metzinger, T. (2003). *Being no one. The self-model theory of subjectivity*. Cambridge, MA: Bradford Books.
- (2007). Empirical perspectives from the self-model theory of subjectivity: A brief summary with examples. *Progress in Brain Research*, 168, 215-278.
[10.1016/S0079-6123\(07\)68018-2](https://doi.org/10.1016/S0079-6123(07)68018-2)
- (2013). Why are dreams interesting for philosophers? The example of minimal phenomenal selfhood, plus an agenda for future research. *Frontiers in Psychology*, 4, 1-17. [10.3389/fpsyg.2013.00746](https://doi.org/10.3389/fpsyg.2013.00746)
- (2014). First-order embodiment, second-order embodiment, third-order embodiment. In L. Shapiro (Ed.) *The Routledge Handbook of Embodied Cognition*. London, UK: Routledge.
- Michal, M., Reuchlein, B., Adler, J., Reiner, I., Beutel, M. E., Vögele, C. & Schulz, A. (2014). Striking discrepancy of anomalous body experiences with normal interoceptive accuracy in depersonalization-derealization disorder. *PLoS One*, 9 (2), e89823-e89823.
[10.1371/journal.pone.0089823](https://doi.org/10.1371/journal.pone.0089823)
- Miller, J. F., Neufang, M., Solway, A., Brandt, A., Trippe, M., Mader, I. & Schulze-Bonhage, A. (2013). Neural activity in human hippocampal formation reveals the spatial context of retrieved memories. *Science*, 342 (6162), 1111-1114.
[10.1126/science.1244056](https://doi.org/10.1126/science.1244056)
- Mittelstaedt, H. (1992). Somatic versus vestibular gravity reception in man. *Annals of the New York Academy of Sciences*, 656, 124-139.
[10.1111/j.1749-6632.1992.tb25204.x](https://doi.org/10.1111/j.1749-6632.1992.tb25204.x)
- (1996). Somatic graviception. *Biological Psychology*, 42 (1-2), 53-74. [10.1016/0301-0511\(95\)05146-5](https://doi.org/10.1016/0301-0511(95)05146-5)
- Moga, M. M., Herbert, H., Hurley, K. M., Yasui, Y., Gray, T. S. & Saper, C. B. (1990). Organization of cortical, basal forebrain, and hypothalamic afferents to the parabrachial nucleus in the rat. *The Journal of Comparative Neurology*, 295 (4), 624-661.
[10.1002/cne.902950408](https://doi.org/10.1002/cne.902950408)
- Mohan, R., Jensen, K. B., Petkova, V. I., Dey, A., Barnsley, N., Ingvar, M. & Ehrsson, H. H. (2012). No pain relief with the rubber hand illusion. *PLoS One*, 7, e52400-e52400. [10.1371/journal.pone.0052400](https://doi.org/10.1371/journal.pone.0052400)
- Moseley, G. L., Olthof, N., Venema, A., Don, S., Wijers, M., Gallace, A. & Spence, C. (2008). Psychologically induced cooling of a specific body part caused by the illusory ownership of an artificial counterpart. *Proceedings of the National Academy of Sciences of the United States of America*, 105 (35), 13169-13171.
[10.1073/pnas.0803768105](https://doi.org/10.1073/pnas.0803768105)
- Nielsen, T. I. (1963). Volition - An new experimental approach. *Scandinavian Journal of Psychology*, 4 (1), 225-230. [10.1111/j.1467-9450.1963.tb01326.x](https://doi.org/10.1111/j.1467-9450.1963.tb01326.x)
- O'Keefe, J. & Conway, D. H. (1978). Hippocampal place units in the freely moving rat: Why they fire where they fire. *Experimental Brain Research*, 31 (4), 573-590. [10.1007/BF00239813](https://doi.org/10.1007/BF00239813)
- O'Keefe, J. & Dostrovsky, J. (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, 34 (1), 171-175. [10.1016/0006-8993\(71\)90358-1](https://doi.org/10.1016/0006-8993(71)90358-1)
- O'Mara, S., Rolls, E. T., Berthoz, A. & Kesner, R. P. (1994). Neurons responding to whole-body motion in the primate hippocampus. *Journal of Neuroscience*, 14 (11), 6511-6523. [10.1.1.64.9583](https://doi.org/10.1.1.64.9583)
- Ödkvist, L. M., Schwarz, D. W. F., Fredrickson, J. M. & Hassler, R. (1974). Projection of the vestibular nerve to the area 3a arm field in the squirrel monkey (*Saimiri sciureus*). *Experimental Brain Research*, 21 (1), 97-105.
[10.1007/BF00234260](https://doi.org/10.1007/BF00234260)

- Ono, T., Nakamura, K., Nishijo, H. & Eifuku, S. (1993). Monkey hippocampal neurons related to spatial and nonspatial functions. *Journal of Neurophysiology*, 70 (4), 1516-1529.
- Ostrowsky, K., Magnin, M., Ryvlin, P., Isnard, J., Guenot, M. & Mauguiere, F. (2002). Representation of pain and somatic sensation in the human insula: A study of responses to direct electrical cortical stimulation. *Cerebral Cortex*, 12 (4), 376-385. [10.1093/cercor/12.4.376](https://doi.org/10.1093/cercor/12.4.376)
- Paillard, J. (1991). Knowing where and knowing how to get there. In J. Paillard (Ed.) *Brain and Space* (pp. 461-481). New York, NY: Oxford University Press.
- Paladino, M. P., Mazzurega, M., Pavani, F. & Schubert, T. W. (2010). Synchronous multisensory stimulation blurs self-other boundaries. *Psychological Science*, 21 (9), 1202-1207. [10.1177/0956797610379234](https://doi.org/10.1177/0956797610379234)
- Parsons, L. M. (1987). Imagined spatial transformation of one's body. *Journal of Experimental Psychology: General*, 116 (2), 172-191. [10.1037/0096-3445.116.2.172](https://doi.org/10.1037/0096-3445.116.2.172)
- Patterson, D. R. & Jensen, M. P. (2003). Hypnosis and clinical pain. *Psychological Bulletin*, 129 (4), 495-521. [10.1037/0033-2909.129.4.495](https://doi.org/10.1037/0033-2909.129.4.495)
- Petkova, V. I. & Ehrsson, H. H. (2008). If I were you: Perceptual illusion of body swapping. *PLoS One*, 3 (12), e3832-e3832. [10.1371/journal.pone.0003832](https://doi.org/10.1371/journal.pone.0003832)
- Petkova, V. I., Bjornsdotter, M., Gentile, G., Jonsson, T., Li, T. Q. & Ehrsson, H. H. (2011). From part- to whole-body ownership in the multisensory brain. *Current Biology*, 21 (13), 1118-1122. [10.1016/j.cub.2011.05.022](https://doi.org/10.1016/j.cub.2011.05.022)
- Pfeiffer, C., Lopez, C., Schmutz, V., Duenas, J. A., Martuzzi, R. & Blanke, O. (2013). Multisensory origin of the subjective first-person perspective: Visual, tactile, and vestibular mechanisms. *PLoS One*, 8 (4), e61751-e61751. [10.1371/journal.pone.0061751](https://doi.org/10.1371/journal.pone.0061751)
- Poucet, B., Lenck-Santini, P. P., Paz-Villagran, V. & Save, E. (2003). Place cells, neocortex and spatial navigation: A short review. *Journal of Physiology*, 97 (4-6), 537-546. [10.1016/j.jphysparis.2004.01.011](https://doi.org/10.1016/j.jphysparis.2004.01.011)
- Péruch, P., Borel, L., Gaunet, F., Thinus-Blanc, G., Magan, J. & Lacour, M. (1999). Spatial performance of unilateral vestibular defective patients in nonvisual versus visual navigation. *Journal of Vestibular Research*, 9 (1), 37-47.
- Ramachandran, V. S., McGeoch, P. D. & Williams, L. (2007). Can vestibular caloric stimulation be used to treat Dejerine-Roussy Syndrome? *Medical Hypotheses*, 69 (3), 486-488. [10.1016/j.mehy.2006.12.036](https://doi.org/10.1016/j.mehy.2006.12.036)
- Ramachandran, V. S., McGeoch, P. D., Williams, L. & Arcilla, G. (2007). Rapid relief of thalamic pain syndrome induced by vestibular caloric stimulation. *Neurocase*, 13 (3), 185-188. [10.1080/13554790701450446](https://doi.org/10.1080/13554790701450446)
- Riva, G., Waterworth, J., Waterworth, E. L. & Mantovani, F. (2011). From intention to action: The role of presence. *New Ideas in Psychology*, 29 (1), 24-37. [10.1016/j.newideapsych.2009.11.002](https://doi.org/10.1016/j.newideapsych.2009.11.002)
- Rizzolatti, G. & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, 27, 169-192. [10.1146/annurev.neuro.27.070203.144230](https://doi.org/10.1146/annurev.neuro.27.070203.144230)
- Rizzolatti, G., Fadiga, L., Gallese, V. & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Brain Research*, 3 (2), 131-141. [10.1016/0926-6410\(95\)00038-0](https://doi.org/10.1016/0926-6410(95)00038-0)
- Robertson, R. G., Rolls, E. T., Georges-François, P. & Panzeri, S. (1999). Head direction cells in the primate pre-subiculum. *Hippocampus*, 9 (3), 206-219. [10.1002/\(SICI\)1098-1063\(1999\)9:3<206::AID-HIPO2>3.0.CO;2-H](https://doi.org/10.1002/(SICI)1098-1063(1999)9:3<206::AID-HIPO2>3.0.CO;2-H)
- Rochat, P. (1998). Self-perception and action in infancy. *Experimental Brain Research*, 123 (1-2), 102-109. [10.1007/s002210050550](https://doi.org/10.1007/s002210050550)
- Röder, C. H., Michal, M., Overbeck, G., van de Ven, V. G. & Linden, D. E. J. (2007). Pain response in depersonalization: A functional imaging study using hypnosis in healthy subjects. *Psychotherapy and Psychosomatics*, 76 (2), 115-121. [10.1159/000097970](https://doi.org/10.1159/000097970)
- Romano, D., Pfeiffer, C., Maravita, A. & Blanke, O. (2014). Illusory self-identification with an avatar reduces arousal responses to painful stimuli. *Behavioural Brain Research*, 261, 275-281. [10.1016/j.bbr.2013.12.049](https://doi.org/10.1016/j.bbr.2013.12.049)
- Roy, J. E. & Cullen, K. E. (2004). Dissociating self-generated from passively applied head motion: Neural mechanisms in the vestibular nuclei. *Journal of Neuroscience*, 24 (9), 2102-2111. [10.1523/JNEUROSCI.3988-03.2004](https://doi.org/10.1523/JNEUROSCI.3988-03.2004)
- Salomon, R., Lim, M., Pfeiffer, C., Gassert, R. & Blanke, O. (2013). Full body illusion is associated with widespread skin temperature reduction. *Frontiers in Behavioral Neuroscience*, 7, 1-11. [10.3389/fnbeh.2013.00065](https://doi.org/10.3389/fnbeh.2013.00065)
- Sanchez-Vives, M. V. & Slater, M. (2005). From presence to consciousness through virtual reality. *Nature Reviews in Neuroscience*, 6 (4), 332-339. [10.1038/nrn1651](https://doi.org/10.1038/nrn1651)
- Sang, F. Y., Jauregui-Renaud, K., Green, D. A., Bronstein, A. M. & Gresty, M. A. (2006). Depersonalisation/derealisation symptoms in vestibular disease. *Journal of Neurology Neurosurgery and Psychiatry*, 77

- (3), 760-766.
[10.1136/jnnp.2007.122119](https://doi.org/10.1136/jnnp.2007.122119)
- Sargolini, F., Fyhn, M., Hafting, T., McNaughton, B. L., Witter, M. P., Moser, M. B. & Moser, E. I. (2006). Conjunctive representation of position, direction, and velocity in entorhinal cortex. *Science*, 312 (5774), 758-762. [10.1126/science.1125572](https://doi.org/10.1126/science.1125572)
- Schilder, P. (1935). *The image and appearance of the human body*. New York, NY: International Universities Press.
- Schlack, A., Sterbing-D'Angelo, S. J., Hartung, K., Hoffmann, K. P. & Bremmer, F. (2005). Multisensory space representations in the macaque ventral intraparietal area. *Journal of Neuroscience*, 25 (18), 4616-4625.
[10.1523/JNEUROSCI.0455-05.2005](https://doi.org/10.1523/JNEUROSCI.0455-05.2005)
- Schneider, R. J., Friedman, D. P. & Mishkin, M. (1993). A modality-specific somatosensory area within the insula of the rhesus monkey. *Brain Research*, 621 (1), 116-120. [10.1016/0006-8993\(93\)90305-7](https://doi.org/10.1016/0006-8993(93)90305-7)
- Schwabe, L., Lenggenhager, B. & Blanke, O. (2009). The timing of temporoparietal and frontal activations during mental own body transformations from different visuospatial perspectives. *Human Brain Mapping*, 30 (6), 1801-1812. [10.1002/hbm.20764](https://doi.org/10.1002/hbm.20764)
- Schwarz, D. W. F. & Fredrickson, J. M. (1971). Rhesus monkey vestibular cortex: A bimodal primary projection field. *Science*, 172 (3980), 280-281.
[10.1126/science.172.3980.280](https://doi.org/10.1126/science.172.3980.280)
- Schwarz, D. W. F., Deecke, L. & Fredrickson, J. M. (1973). Cortical projection of group I muscle afferents to areas 2, 3a, and the vestibular field in the rhesus monkey. *Experimental Brain Research*, 17 (5), 516-526.
[10.1007/BF00234865](https://doi.org/10.1007/BF00234865)
- Serino, A., Alsmith, A., Costantini, M., Mandrigin, A., Tajadura-Jimenez, A. & Lopez, C. (2013). Bodily ownership and self-location: Components of bodily self-consciousness. *Consciousness and Cognition*, 22 (4), 1239-1252. [10.1016/j.concog.2013.08.013](https://doi.org/10.1016/j.concog.2013.08.013)
- Serino, A., Giovagnoli, G. & Ladavas, E. (2009). I feel what you feel if you are similar to me. *PLoS One*, 4 (3), e4930-e4930. [10.1371/journal.pone.0004930](https://doi.org/10.1371/journal.pone.0004930)
- Serino, A., Pizzoferrato, F. & Ladavas, E. (2008). Viewing a face (especially one's own face) being touched enhances tactile perception on the face. *Psychological Science*, 19 (5), 434-438.
[10.1111/j.1467-9280.2008.02105.x](https://doi.org/10.1111/j.1467-9280.2008.02105.x)
- Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends of Cognitive Sciences*, 17 (11), 565-573. [10.1016/j.tics.2013.09.007](https://doi.org/10.1016/j.tics.2013.09.007)
- Smith, A. J. T. (2010). Comment: Minimal conditions for the simplest form of self-consciousness. In H. C. Fuchs, C. Sattel & P. Henningsen (Eds.) *The embodied self: Dimensions, coherence, disorders*. Stuttgart, GER: Schattauer.
- Stackman, R. W. & Taube, J. S. (1997). Firing properties of head direction cells in the rat anterior thalamic nucleus: Dependence on vestibular input. *Journal of Neuroscience*, 17 (11), 4349-4358.
- Stackman, R. W., Clark, A. S. & Taube, J. S. (2002). Hippocampal spatial representations require vestibular input. *Hippocampus*, 12 (3), 291-303.
[10.1002/hipo.1112](https://doi.org/10.1002/hipo.1112)
- Stratton, G. M. (1899). The spatial harmony of touch and sight. *Mind*, 8 (32), 492-505.
- Sugiuchi, Y., Izawa, Y., Ebata, S. & Shinoda, Y. (2005). Vestibular cortical area in the periarculate cortex: Its afferent and efferent projections. *Annals of the New York Academy of Sciences*, 1039 (2005), 111-123.
[10.1196/annals.1325.011](https://doi.org/10.1196/annals.1325.011)
- Suzuki, K., Garfinkel, S. N., Critchley, H. D. & Seth, A. K. (2013). Multisensory integration across exteroceptive and interoceptive domains modulates self-experience in the rubber-hand illusion. *Neuropsychologia*, 51 (13), 2909-2917.
[10.1016/j.neuropsychologia.2013.08.014](https://doi.org/10.1016/j.neuropsychologia.2013.08.014)
- Suzuki, M., Kitano, H., Ito, R., Kitanishi, T., Yazawa, Y., Ogawa, T. & Kitajima, K. (2001). Cortical and subcortical vestibular response to caloric stimulation detected by functional magnetic resonance imaging. *Cognitive Brain Research*, 12 (3), 441-449.
[10.1016/S0926-6410\(01\)00080-5](https://doi.org/10.1016/S0926-6410(01)00080-5)
- Tajadura-Jiménez, A., Longo, M., Coleman, R. & Tsakiris, M. (2012). The person in the mirror: Using the enfacement illusion to investigate the experiential structure of self-identification. *Consciousness and Cognition*, 21 (4), 1725-1738.
[doi:10.1016/j.concog.2012.10.004](https://doi.org/10.1016/j.concog.2012.10.004)
- Taube, J. S. (2007). The head direction signal: Origins and sensory-motor integration. *Annual Review of Neuroscience*, 30, 181-207.
[10.1146/annurev.neuro.29.051605.112854](https://doi.org/10.1146/annurev.neuro.29.051605.112854)
- Tomlinson, R. D. & Robinson, D. A. (1984). Signals in vestibular nucleus mediating vertical eye movements in the monkey. *Journal of Neurophysiology*, 51 (6), 1121-1136.
- Tsakiris, M. (2010). My body in the brain: A neurocognitive model of body-ownership. *Neuropsychologia*, 48 (3), 703-712. [10.1016/j.neuropsychologia.2009.09.034](https://doi.org/10.1016/j.neuropsychologia.2009.09.034)

- Tsakiris, M., Hesse, M. D., Boy, C., Haggard, P. & Fink, G. R. (2007). Neural signatures of body ownership: A sensory network for bodily self-consciousness. *Cerebral Cortex*, 17 (10), 2235-2244. [10.1093/cercor/bhl131](https://doi.org/10.1093/cercor/bhl131)
- Tsakiris, M., Prabhu, G. & Haggard, P. (2006). Having a body versus moving your body: How agency structures body-ownership. *Consciousness and Cognition*, 15 (2), 423-432. [10.1016/j.concog.2005.09.004](https://doi.org/10.1016/j.concog.2005.09.004)
- Tsakiris, M., Schutz-Bosbach, S. & Gallagher, S. (2007). On agency and body-ownership: Phenomenological and neurocognitive reflections. *Consciousness and Cognition*, 16 (3), 645-660. [10.1016/j.concog.2007.05.012](https://doi.org/10.1016/j.concog.2007.05.012)
- Tsakiris, M., Tajadura-Jimenez, A. & Costantini, M. (2011). Just a heartbeat away from one's body: Interoceptive sensitivity predicts malleability of body-representations. *Proceedings of the Royal Society B: Biological Sciences*, 278 (1717), 2470-2476. [10.1098/rspb.2010.2547](https://doi.org/10.1098/rspb.2010.2547)
- Ulanovsky, N. & Moss, C. F. (2007). Hippocampal cellular and network activity in freely moving echolocating bats. *Nature Neuroscience*, 10 (2), 224-233. [10.1038/nm1829](https://doi.org/10.1038/nm1829)
- Väljamäe, A. (2009). Auditorily-induced illusory self-motion: A review. *Brain Research Reviews*, 61 (2), 240-255. [10.1016/j.brainresrev.2009.07.001](https://doi.org/10.1016/j.brainresrev.2009.07.001)
- Vaitl, D., Mittelstaedt, H., Saborowski, R., Stark, R. & Baisch, F. (2002). Shifts in blood volume alter the perception of posture: Further evidence for somatic graviception. *International Journal of Psychophysiology*, 44 (1), 1-11. [10.1016/S0167-8760\(01\)00184-2](https://doi.org/10.1016/S0167-8760(01)00184-2)
- Vallar, G., Sterzi, R., Bottini, G., Cappa, S. & Rusconi, M. L. (1990). Temporary remission of left hemianesthesia after vestibular stimulation. A sensory neglect phenomenon. *Cortex*, 26 (1), 123-131. [10.1016/S0010-9452\(13\)80078-0](https://doi.org/10.1016/S0010-9452(13)80078-0)
- Van Elk, M. & Blanke, O. (2014). Imagined own-body transformations during passive self-motion. *Psychological Research*, 78 (1), 18-27. [10.1007/s00426-013-0486-8](https://doi.org/10.1007/s00426-013-0486-8)
- Van Elk, M., Lenggenhager, B., Heydrich, L. & Blanke, O. (2014). Suppression of the auditory N1-component for heartbeat-related sounds reflects interoceptive predictive coding. *Biological Psychology*, 99, 172-182. [10.1016/j.biopsycho.2014.03.004](https://doi.org/10.1016/j.biopsycho.2014.03.004)
- van Essen, D. C., Drury, H. A., Dickson, J., Harwell, J., Hanlon, D. & Anderson, C. H. (2001). An integrated software suite for surface-based analyses of cerebral cortex. *Journal of the American Medical Informatics Association*, 8 (5), 443-459. [10.1098/rspb.2010.2547](https://doi.org/10.1098/rspb.2010.2547)
- Vogele, K. & Fink, G. R. (2003). Neural correlates of the first-person-perspective. *Trends in Cognitive Science*, 7 (1), 38-42. [10.1016/S1364-6613\(02\)00003-7](https://doi.org/10.1016/S1364-6613(02)00003-7)
- Waespe, W. & Henn, V. (1978). Conflicting visual-vestibular stimulation and vestibular nucleus activity in alert monkeys. *Experimental Brain Research*, 33 (2), 203-211. [10.1007/BF00238060](https://doi.org/10.1007/BF00238060)
- Werner, N. S., Mannhart, T., Reyes Del Paso, G. A. & Duschek, S. (2014). Attention interference for emotional stimuli in cardiac interoceptive awareness. *Psychophysiology*, 51 (6), 573-578. [10.1111/psyp.12200](https://doi.org/10.1111/psyp.12200)
- Wicker, B., Keysers, C., Plailly, J., Royet, J. P., Gallese, V. & Rizzolatti, G. (2003). Both of us disgusted in my insula: The common neural basis of seeing and feeling disgust. *Neuron*, 40 (3), 655-664. [10.1016/S0896-6273\(03\)00679-2](https://doi.org/10.1016/S0896-6273(03)00679-2)
- Wiener, S. I., Berthoz, A. & Zugaro, M. B. (2002). Multi-sensory processing in the elaboration of place and head direction responses by limbic system neurons. *Brain Research Cognitive Brain Research*, 14 (1), 75-90. [10.1016/S0926-6410\(02\)00062-9](https://doi.org/10.1016/S0926-6410(02)00062-9)
- Windt, J. M. (2010). The immersive spatiotemporal hallucination model of dreaming. *Phenomenology and the Cognitive Sciences*, 9 (2), 295-316. [10.1007/s11097-010-9163-1](https://doi.org/10.1007/s11097-010-9163-1)
- Wissmath, B., Weibel, D., Schmutz, J. & Mast, F. W. (2011). Being present in more than one place at a time? Patterns of mental self-localization. *Consciousness and Cognition*, 20 (4), 1808-1815. [10.1016/j.concog.2011.05.008](https://doi.org/10.1016/j.concog.2011.05.008)
- Wolpert, D. M. & Miall, R. C. (1996). Forward models for physiological motor control. *Neural Networks*, 9 (8), 1265-1279. [10.1016/S0893-6080\(96\)00035-4](https://doi.org/10.1016/S0893-6080(96)00035-4)
- Yartsev, M. M., Witter, M. P. & Ulanovsky, N. (2011). Grid cells without theta oscillations in the entorhinal cortex of bats. *Nature*, 479 (7371), 103-107. [10.1038/nature10583](https://doi.org/10.1038/nature10583)
- Yates, B. J. & Bronstein, A. M. (2005). The effects of vestibular system lesions on autonomic regulation: Observations, mechanisms, and clinical implications. *Journal of Vestibular Research*, 15 (3), 119-129.
- Zennou-Azogui, Y., Bourgeon, S. & Xerri, C. (2011). Hypergravity experience during development alters forepaw somatosensory maps and influences cortical experience-dependent plasticity in adult rat. *Presented at the Conference on Development and Plasticity of Thalamocortical Systems*. Arolla, SUI.
- zu Eulenburg, P., Baumgärtner, U., Treede, R. D. & Dieterich, M. (2013). Interoceptive and multimodal func-

- tions of the operculo-insular cortex: Tactile, nociceptive and vestibular representations. *NeuroImage*, 83, 75-86. [10.1016/j.neuroimage.2013.06.057](https://doi.org/10.1016/j.neuroimage.2013.06.057)
- zu Eulenburg, P., Caspers, S., Roski, C. & Eickhoff, S. B. (2012). Meta-analytical definition and functional connectivity of the human vestibular cortex. *NeuroImage*, 60 (1), 162-169. [10.1016/j.neuroimage.2011.12.032](https://doi.org/10.1016/j.neuroimage.2011.12.032)

Perspectival Structure and Vestibular Processing

A Commentary on Bigna Lenggenhager & Christophe Lopez

Adrian Alsmith

I begin by contrasting a taxonomic approach to the vestibular system with the structural approach I take in the bulk of this commentary. I provide an analysis of perspectival structure. Employing that analysis and following the structural approach, I propose three lines of empirical investigation to selectively manipulate and measure vestibular processing and perspectival structure. The hope is that this serves to indicate how interdisciplinary research on vestibular processing might advance our understanding of the structural features of conscious experience.

Keywords

Egocentric | Egomotion | First-person perspective | Galvanic vestibular stimulation (GVS) | GVS | Head-mounted display | Perspective | Phenomenal groove | Phenomenal grooves | Scalp EEG | Self-consciousness | Structural features of consciousness | Tendon vibration stimulation | The body-swap illusion | The full-body illusion | The senses | Vestibular | Vestibular evoked potentials

Commentator

[Adrian Alsmith](#)
adrianjalsmith@gmail.com
Københavns Universitet
Copenhagen, Denmark

Target Authors

[Bigna Lenggenhager](#)
bigna.lenggenhager@usz.ch
University Hospital
Zurich, Switzerland

[Christophe Lopez](#)
christophe.lopez@univ-amu.fr
Aix Marseille Université
Marseille, France

Editors

[Thomas Metzinger](#)
metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Structural vs. taxonomic approaches to vestibular processes

Philosophical work on the senses has largely been concerned with taxonomic issues: What makes an event sensory? Under which sensory kind should that event be classified? Answering these questions requires criteria of individuation. These would enable us to determine

whether an event is the same as (or different to) sensory events in general and whether it is the same as (or different to) sensory events of a specific kind. A criterion of the first sort would allow us to identify vestibular events as sensory events. This would justify the belief that vesti-

bular processes are sensory processes. A criterion of the second sort would allow us to identify vestibular sensory events as being of a specific kind, i.e., distinctively vestibular sensory events. This would justify the belief that there is such a thing as a vestibular sense. Failing to provide a criterion of the first sort would force one to classify vestibular events as non-sensory. But even if one were able to determine that vestibular events are sensory, one would still require a criterion of the second sort to classify vestibular events as sensory events of a kind that is distinct from, e.g., visual or haptic events.

To expand on this last point: as Lenggenhager and Lopez so masterfully describe, central vestibular processes are inherently multisensory, and as a consequence there is scarcely a part of our sensory and cognitive life that vestibular processes leave untouched (see especially §2.2 of the target article). But then, if vestibular processes are implicated in so many sensory and cognitive processes, it may be most accurate to see vestibular processing as simply a common part of many processes, rather than as an independent sensory system. That is, one may begin to seriously consider the possibility that vestibular processing does not constitute a form of sensory processing of its own kind, but rather constitutes a form of processing common to various other processes that are themselves sensory. This is, in effect, an issue that arises from applying a criterion for individuating the senses that includes the physiology (and neurophysiology) of the entire system. One might not be forced to this conclusion if one used an alternative criterion (Macpherson 2011a, 2011b). But it seems that each of the criteria commonly discussed would generate their own problems. For instance, employing a more restrictive criterion that delimited sensory systems according to their peripheral sensory organs would face the issue of whether the sensory organs of the vestibular system ought to include or exclude the so-called “truncal” or “somatic” graviceptors (Mittelstaedt 1992, 1996; Vaitl et al. 2002). Similar issues would be faced when attempting to individuate the senses in terms of a distinctive proximal stimulus. Alternatively, one might

individuate the senses by means of certain distinctive experiences: vision distinctively represents the brightness, hue, and saturation of colours; audition represents the volume, pitch, and tone of sounds. The natural candidates for the vestibular system would be experiences that represent verticality, rotation, and translation. But whilst it is certain that the vestibular system typically contributes to experiences of verticality, rotation, and translation, these are all experiences of a kind that can be had through visual sensation alone, or through a combination of visual, somatic, and proprioceptive sensation. Moreover, although vertiginous experiences are the hallmark of vestibular dysfunction, these are either experiences of rotation, which brings us back to the aforementioned issue, or they are more vaguely classified as pseudo-vertiginous experiences of dizziness that may have any number of non-vestibular aetiologies. Suffice to say that it may be surprisingly difficult to find appropriate criteria to justify the claim that there is such a thing as a *distinctively* vestibular sensory process.

The foregoing characterises what would be the typical philosophical approach to the vestibular system, *qua* sensory system. This *taxonomic approach* captures certain philosophical interests, but it is completely inadequate for the task of bringing out the significance of the scope of the vestibular system’s influence. An alternative, *structural approach* focuses on the role played by vestibular events in processes that exhibit a certain kind of structure, to determine the contribution of those events to that structure. Note that the structural and taxonomic approaches are independent, insofar as they have different epistemic goals. They aim to further our knowledge in different ways. The goal of the taxonomic approach is to determine whether, and if so why, there is a distinctive sensory system of a certain kind. The goal of the structural approach is to determine whether, and if so how, a certain kind of process contributes to a certain kind of structure. By assuming that one can identify processes as objects of study without first employing an exhaustive taxonomy, a structural approach can assume that there are such things as vestibular

processes without any commitment to these processes being wholly distinct from others. And by tracking the varied yet systematic effects of vestibular processes, one can determine whether vestibular processes contribute to a certain kind of structure, irrespective of, whether or not the vestibular system is a distinctive sensory system. As vestibular processes are implicated in so many and various sensory and cognitive processes, the structural approach seems to be the most fruitful in terms of the amount we might learn. It also seems more fruitful in terms of the kind of knowledge we might gain. For we may learn nothing about how vestibular processes affect our experiential life by learning that vestibular processes may not be, in the final analysis, of a distinctive sensory kind. But we will certainly learn something about how vestibular processes affect our experiential life by learning that vestibular processes contribute to a certain experiential structure. Accordingly, I leave aside taxonomic issues in the rest of this commentary and focus on structural issues. Specifically, I focus on issues concerning the role of the vestibular system in providing a particular kind of structure to our experience of the body and the world, namely a perspectival structure.

To begin with, we need a preliminary analysis of experiential phenomena that exhibit perspectival structure. I will call these *perspectival phenomena*. In the next section, I offer a rudimentary analysis of perspectival structure, the aim of which is to show that perspectival phenomena are more differentiated than commonly recognised. In the following three sections, I propose three lines of empirical investigation. Each would attempt to selectively study perspectival phenomena through measurement and manipulation of vestibular processes. If the experiments proposed yielded interesting results, they would further our knowledge of how vestibular processes affect the perspectival structure of our experiential life. Accordingly, the overall aim is to demonstrate how an analysis of perspectival structure might fruitfully interface with empirical research and facilitate understanding of structural features of conscious experience that would otherwise be obscured.

2 The differentiation of perspectival phenomena

The notion of a subjective perspective (sometimes described as a *first-person perspective*) is at the core of contemporary research on bodily self-consciousness (Blanke & Metzinger 2009; Metzinger 2003, 2009). However, its role has often been merely facilitative, serving as a means to study *other* components of bodily self-consciousness, such as the experience of bodily agency, ownership, and self-location (Ehrsson 2007; Lenggenhager et al. 2007; Petkova et al. 2011a, see Serino et al. 2013 for review). Consequently, the fact that the very notion of perspective covers a range of distinct phenomena has tended to be overlooked.¹ Referring to someone's perceptual experience as having a perspectival structure may mean any one of several distinct things. It may mean that there is an *origin* to her sensory field, relative to which certain things (or parts of things) are perceptible and perceived from a particular direction and relative to which certain other things (or parts of things) are not perceptible or noticeably occluded.² Alternatively, it may mean that her experience is organised according to an *egocentric* frame of reference centred upon her body, according to which she experiences locations as situated relative to a particular point at the intersection of three orthogonal axes. Or it may be that, thanks to *egomotion*, the flow of her sensory experience is such that she can see where she is headed as she moves. Taking *another individual's* perspective into account in social interactions can involve either of the first two forms of perspective (Moll & Meltzoff 2011).

¹ My discussion is restricted to spatial perspectival phenomena; I omit discussion of the respects in which temporal experience may be perspectival. This is mostly for the sake of simplicity. However, there is good reason to think that we represent time in a manner that is asymmetrically dependent upon the ways in which we represent space (Boroditsky 2000; Casasanto & Boroditsky 2008). Addressing issues concerning the structure of spatial experience first may thus be prudent.

² This notion is intended to capture the idea that there is a point of "origin" to the so-called line of sight (which is not so much a line as an angle). This corresponds to perhaps the earliest documented notion of perceptual perspective, associated with what Euclid and Ptolemy respectively called the "visual pyramid" and "visual cone", where the apex (origin) of the pyramid or cone is at the eye and the base at the object (Howard 2012).

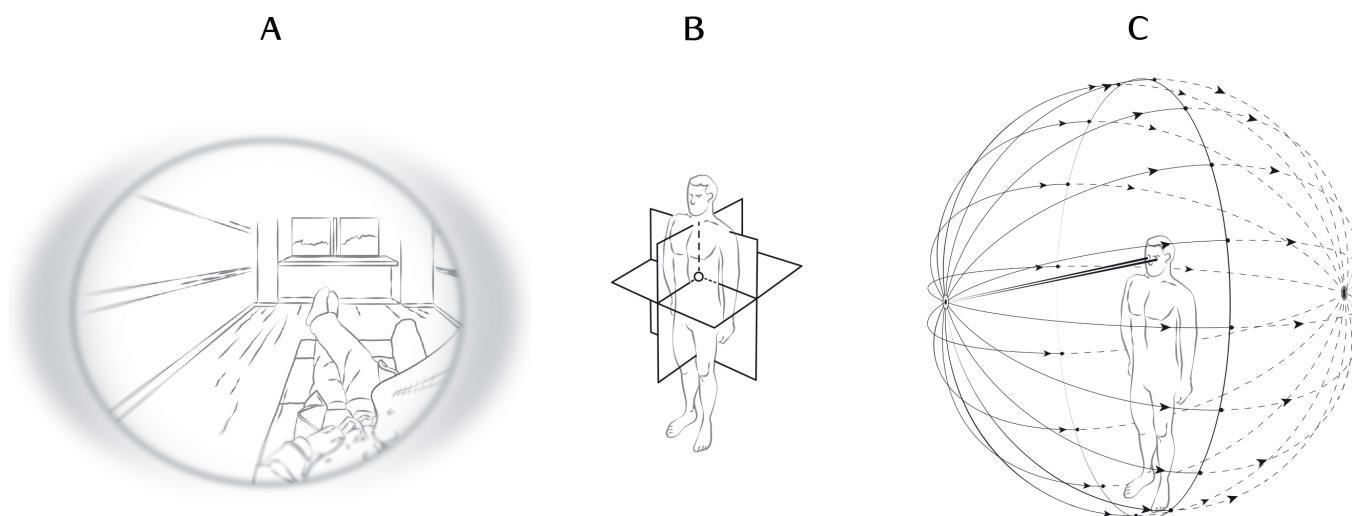


Figure 1: Three forms of perspectival structure. **A.** An artistic rendition of a human monocular visual field. After Mach 1959, p.18. **B.** An egocentric frame of reference centred upon a human torso. **C.** The directions of deformations in the visual field specifying egomotion. Cf. Gibson 1950, p. 123.

Moreover, the perspective of the subject need not figure explicitly in the experience for it to be perspectival; perspective can structure perceptual experience implicitly, by determining the way in which objects are experienced, without itself being part of the content of the *experience* (Campbell 1994; Merleau-Ponty 2002; Perry 1993; Zahavi 2005).

We can summarise these remarks by saying that perspectival phenomena in spatial experience vary along three dimensions.³ First, perspectival structure can take at least three forms:⁴

- Origin of a sensory field (*origin*)
- Centre of an egocentric frame of reference (*egocentric*)
- Focal point of a sensory flow field in action (*egomotion*)

³ I do not intend the following to be exhaustive. Moreover, although all of the perspectival phenomena that I discuss are visual, I do believe that each of the forms of perspectival structure that I describe also characterises perspectival experience in haptic perception.

⁴ The most I intend to claim here is that these forms of perspectival structure are non-identical. Perhaps the origin of a given sensory field, the centre of a given egocentric frame of reference, and the focal point of a given sensory flow field could occupy the same location under some description. However, this certainly need not always be the case. Moreover, each form of perspectival structure will present the objects of perceptual experience as related to the subject of experience in different ways, e.g., as only partially visible, as straight ahead, or as in one's way. Below I will suggest various ways in which these might be selectively manipulated, but I do not intend to make the case that forms of perspectival structure can be dissociated from one another.

Perspectival phenomena that exhibit any of these forms of perspectival structure can vary along two further dimensions: the perspective of a given perspectival experience may be either *implicit* or *explicit*, and may be attributed to the subject or to another individual. A perspective is *explicit* in a perspectival experience if the subject is consciously aware of the location of the origin, centre, or focal point in question; it is *implicit* if the subject is not.⁵ The perspective in question may belong to the subject, a *first-person perspective*, or it may belong to another individual, a *third-person perspective*.

This simple framework enables one to study perspectival phenomena selectively, rather than studying an undifferentiated cluster of perspectival phenomena simultaneously. In the sections that follow, I shall suggest a number of ways in which one might engage in such a selective study of perspectival phenomena by in-

⁵ When a perspective is explicit, the location of the origin, centre, or focal point is part of the content of the experience. Any beliefs that the subject has about the location in question do not go beyond the content of that experience (cf. Peacocke 1999, p. 265). The experience may represent the location in question in an imprecise or wholly incorrect manner; the subject's beliefs will be correspondingly imprecise or incorrect. Implicit perspectives structure experience without being part of the content of experience. I leave it open whether implicit perspectives are nevertheless experienced *qua* structural feature, or whether, for example, they are merely formal structures that determine the ways in which things are experienced, without themselves being experienced. Issues like this are difficult to evaluate, but for discussion see Alsmith (2012).

tervening upon and registering the activation of vestibular processes.

3 Perspectival variation in multisensory stimulation

One consequence of not distinguishing between perspectival phenomena is that the notion of a *first-person perspective* becomes ambiguous. One can clearly see this ambiguity in descriptions of the role of *first-person perspective* in the multisensory stimulation protocols developed in recent work on the neuroscience of bodily self-consciousness. These protocols all involve participants being touched on their torso whilst visually observing a body-shape (either the body of another person, a mannequin, or a virtual body) being touched on its torso. The protocols differ along two dimensions: the side of the torso stimulated and the location of the origin of the participants' line of sight with respect to the body being observed. In one protocol, the body-swap illusion, participants are stroked on their chest whilst they look at a body being stroked on its chest from a position located where its head would be (cf. Ehrsson 2007; see Petkova et al. 2011b; Petkova & Ehrsson 2008; Petkova et al. 2011a). In another protocol, the full-body illusion, participants are stroked on their back, whilst they observe a body from behind being stroked on its back from a position entirely removed from its location (Ionta et al. 2011; Lenggenhager et al. 2007; Pfeiffer et al. 2013). The body-swap illusion protocol is often distinguished from the full-body illusion protocol as involving first-person perspective as an independent variable (Petkova et al. 2011a). However, recent work on the full-body illusion has demonstrated effects that the authors describe as changes in first-person perspective (Pfeiffer et al. 2014): Participants lain prone whilst feeling and observing strokes on the back report experiences of either looking up or down at the body they observe (Ionta et al. 2011). These variations in report seem to depend upon the individual's relative weighting of vestibular and visual gravitational cues (Pfeiffer et al. 2013).

Admitting the differentiation of perspectival phenomena allows us to make sense of the differences in use of the term *first-person*

perspective. In the terms introduced in the previous section, the *first-person perspective* in the body-swap illusion is an *origin* perspective. It presents the typical view of one's own body with a line of sight originating in the head. The *first-person perspective* in the full-body illusion is an *egocentric* perspective. It forms the centre of an egocentric frame of reference, according to which the observed body occupies a location in a particular egocentric direction (up or down). Distinguishing these forms of first-person perspectival experience reveals that each of these protocols facilitates manipulation of a distinct form of perspectival experience. It also sheds light on the fact that the differences in vestibular and somatosensory processing between these forms of perspectival experience have yet to be compared.

One way of conducting such a comparison would be to use virtual reality display techniques to present an individual with two avatars in series, whilst measuring time-locked vestibular evoked potentials via scalp EEG.

Experiment 1: Participants are stroked on *both* their chest *and* their back whilst supine, whilst wearing a head-mounted display. In the meantime, participants observe either the chest of Avatar 1 being stroked on its chest, presented from a position corresponding to the avatar's head, as in the body-swap illusion, or they observe Avatar 2 being stroked on its back, as in the full-body illusion. Ideally, the two avatars are presented in the same viewing, such that the participant views one avatar and then in a continuous movement shifts their gaze to view the other.⁶

I have claimed that each of the two protocols conjoined in this proposed experiment facilitates manipulation of different forms of perspectival experience. If this is correct, then finding significant differences in vestibularly-evoked potentials between observation of Avatar 1 and Avatar 2 would be a first step in determining differences in vestibular processing between these forms of perspectival experience.

⁶ This would be, I take it, as close as practically possible to viewing the two avatars at the same time, given limitations in the field of view.

As noted earlier, there do seem to be individual differences in the contents of *egocentric* perspectival experience in the full-body illusion. This would suggest that some individuals, those who are more heavily dependent upon vestibular gravitational cues to determine orientation, would experience themselves as looking upwards at Avatar 2. Whereas if the right *visual* gravitational cues were provided, some individuals may experience themselves as looking downwards at Avatar 2 (Ionta et al. 2011; Pfeiffer et al. 2013). This might allow the investigation of the relationship between *egocentric* perspectives and *egomotion* perspectives, by incorporating a second phase into a new experiment:

Experiment 2: Phase 1: experiment 1, described above. Phase 2: Participants continue to be stroked on their back and chest. Participants fixate upon Avatar 2 and observe it rotating about a horizontal axis, whilst being visibly stroked on its back and chest. Both reports of experienced orientation (upward vs. downward) and reports of experienced *egomotion* are gathered.

Participants may experience themselves as rotating around a horizontal axis in just the way they observe Avatar 2 rotating. Alternatively, they might experience themselves as revolving around Avatar 2. In particular, what would be of interest would be the way in which any resultant illusory experiences of *egomotion* might correlate with experienced *egocentric* orientation (upward vs. downward). Moreover, individual differences in experienced *egocentric* orientation might even predict the contents of experienced *egomotion*. This would be a major step in determining both the relative influence of vestibular processing on these forms of perspectival experience and the relationship between these forms of perspectival experience.

4 Perspectival variation in misalignment

In much recent philosophical and neuroscientific research on self-consciousness, the experienced *first-person perspective* is treated as a simple phenomenon identified with the experienced origin of an *egocentric* frame of reference centred upon an individual's own body (Blanke & Metzinger 2009; Vogeley & Fink 2003). But *egocentric* perspective, despite being an *apparently* simple phenomenon, is in fact as potentially complex as the macroscopic structure of the body itself (Smith 2010). Human bodies are composed of a number of parts that are to some degree independently mobile, any of which may serve to centre a distinct *egocentric* frame of reference. As this observation is well known, we may presume that theorists who treat *egocentric* perspective as simple are assuming that locations in these various *egocentric* frames of reference are translated into a single, *ultimate* *egocentric* frame reference which itself determines *egocentric* perspectival phenomena.

However, neurophysiological and neuropsychological research on spatial representation suggests independent motivation for this ultimate frame being centred upon the head (e.g., Avillac et al. 2005) or the torso (e.g., Karnath et al. 1991). By rotating head and torso in opposite directions, an *egocentric* frame of reference centred upon the head can be misaligned with another frame centred upon the torso. In such a “misalignment” situation, a single object may be “to the right” with respect to the head and “to the left” with respect to the torso (Longo & Alsmith 2013). Following Christopher Peacocke's (1992) description of the phenomenology of experienced direction, one would hypothesise that differences in experienced posture would determine differences in *egocentric* perspectival experience.⁷ One could thus use mis-

alignment to investigate the relationship between experienced orientation and experienced *egomotion*. This would be a major step in determining both the relative influence of vestibular processing on these forms of perspectival experience and the relationship between these forms of perspectival experience.

⁷ Peacocke writes: “The use of a particular set of labeled axes in giving part of the content of an experience is not a purely notational or conventional matter. The appropriate set of labeled axes captures distinctions in the phenomenology of experience itself. Looking straight ahead at Buckingham Palace is one experience. It is another to look at the palace with one's face still toward it but with one's body turned toward a point on the right. In this second case the palace is experienced as being off to one side from the direction of straight ahead, even if the view remains exactly the same as in the first case” (1992, p. 62). Assuming that Peacocke's prediction is correct, then in this example changes in the *egocentric* perspectival structure of visual experience follow changes in the orientation of the torso. By misaligning the torso from the direction of the gaze, one discerns that (in the case as described) the appropriate set of labeled axes centre upon the torso. In the paradigm described in experiment 3, both head and torso may be misaligned with the individual's gaze. This makes it possible to determine the contribution of both head- and torso-centred frames of reference to the individual's *egocentric* perspectival experience of a given location. It would then be possible to discern whether, for the *egocentric* perspectival experience of a given location: (i) the appropriate set of axes centre on the torso; (ii) the axes centre on the head; (iii) both sets of axes make relative contributions to the structure of the experience.

alignment situations to determine the respective contributions of the head and the torso to the organisation of *egocentric* perspectival experience at a given point in time in the following experiment:

Experiment 3: Standing with their head and torso aligned or misaligned $\pm 15^\circ$, participants perform a task that involves either an *explicit* or only an *implicit egocentric* perspective (see below). The angular deviation of the stimulus in relation to the head and/or torso is recorded, such that one would be able to assess the respective contributions of each body-part's orientation to the participants' *egocentric* perspectival judgments. Participants would receive either galvanic vestibular stimulation (GVS) or tendon vibration stimulation to precisely assess the relative contribution of vestibular processes to *egocentric* perspective.

In more detail, the suggestions are these. For an explicit task, stimuli could be presented across the entire visual field in regular intervals, varying in distance and elevation, and participants would judge whether a stimulus presented looks "to their left or to their right". A potential limitation of the explicit task is that in using overt left/right judgements, participants' responses may reflect a stipulated meaning of these terms that is independent of the *egocentric* perspectival structure of their experience. However, a recent study using a covert attentional cuing paradigm found that rotation of the torso primes participants to respond more quickly to visual stimuli appearing on the side of a computer screen congruent to the direction of rotation (Grubb & Reed 2002).⁸ One could adapt this paradigm to directly compare the respective influences of head and torso by rotating the head and/or the torso $\pm 15^\circ$ relative to the screen where stimuli would be presented. Target and cuing visual stimuli would appear on either congruent or incongruent sides of the screen and participants would make speed responses to indicate whether the target appears to the left or the right on each trial. Again, as

the angular deviation of the stimulus in relation to the head and/or torso would be known, one would be able to assess the respective contributions of each body-part's orientation to the participant's *egocentric* perspectival judgments.

Based on previous work, I would expect participants' judgements to implicate *both* their head *and* torso as determining their *egocentric* perspectival experience (Alsmith & Longo 2014). More specifically, I would expect that both head- and torso-centred reference frames would influence *explicit* and *implicit egocentric* perspectival phenomena (Longo & Alsmith 2013), though the exact weighting will be unequal at lateral extremes of each body part and will differ between individuals (Alsmith et al. in preparation). The further prediction would be that manipulating vestibular and proprioceptive processing will modulate felt postural misalignment and thereby systematically influence performance on *explicit* and *implicit egocentric* perspectival tasks.

5 Perspectival variation in sensorimotor control

Arguably, one of the core structural features of the experience of intentionally-directed bodily movement is the presentation of the agent as the "perspectival source" of the motion experienced (Horgan et al. 2003; Marcel 2006). However, a strikingly robust experimental finding is that individuals will correct for a deviation introduced into a movement they perform via a bias in visual input, thereby ensuring the action they intend achieves its goal, whilst nevertheless *not* reporting such corrections in their movement (Fournier & Jeannerod 1998; Knoblich & Kircher 2004; Slachevsky et al. 2001). Recent developments of this paradigm have adapted it to test *explicit egomotion* perspectival experience in walking movements, by using a motion-tracked avatar, observed from the rear. Kannape and colleagues found that by introducing a slight bias into the subject's visual experience of the trajectory of the avatar, they could induce subjects to perform appropriate corrective movements in walking to a target, whilst not noting the discrepancy between their actual movements and the avatar (Kannape

⁸ It is perhaps worth noting that by "congruent" I intend the more general sense of the term, as often used in describing the design of behavioural studies, the meaning of which is equivalent to "in agreement". I do not intend the more specific geometrical sense of the term, which expresses identity of a certain kind, typically of form.

et al. 2010). Again, the corrections went largely unnoticed within a certain range of angular deviation between observed and actual movements.⁹ Thus, a natural explanation of the pattern of data is that the mechanisms enabling the experience of agency present bodily movements in a manner that is far more coarse-grained than the level of detail required to make corrective changes in movement trajectory. In short, *egomotion* perspectives structure experiences of intentionally-directed bodily movement. They do so by specifying what we might call coarse-grained phenomenal grooves, within which a movement must unfold if it is to seem like the movement that the subject intended or is trying to perform.

Strangely, as yet the potential contributions of the vestibular system to the structuring of agentive experience by *egomotion perspective* have not been manipulated. Moreover, as noted, the work that has been done in this area has been restricted to *explicit egomotion* perspectival phenomena. A natural further step would be to investigate the nature of vestibular processing in implicit *egomotion* perspective, by controlling a participant's optic flow in a manner corresponding to the control of the avatar's motion in Kannape and colleagues' original study.

Experiment 4: Study 1: Participants view a textured environment via HMD in which optical flow fields are regulated by their motion-tracked movements. Study 2: Participants control a motion-tracked, real-time avatar seen from behind. In both studies, participants are tasked with walking directly towards a virtual target. All the while, they either receive GVS or sham stimulation and visual feedback (optic flow or avatar position) that is either faithful to motion-tracking or systematically deviated left/right of the participant's mid-line, as a function of distance from a point of displacement onset.

Participant trajectory could thus be compared to the dynamics of the flow field or avatar trajectory and participants could be

asked to rate the degree to which their movements in the virtual environment or the movements of the avatar corresponded to their actual movements, as respective measures of *implicit* and explicit *egomotion* perspectival experience. The question would be whether, in trials in which GVS is applied, the range of angular deviation in which participants would judge that movements in the virtual environment correspond to their own would be equal to or larger than trials in which participants receive only biased visual feedback. If the latter occurs, then in the evocative terms used above, it would suggest that vestibular processes are one of the determinants of the coarseness of the phenomenal groove specified by an *egomotion* perspective.

6 Conclusion

I began by contrasting a taxonomic approach to the vestibular system with the structural approach I have taken in the bulk of this commentary. I then provided an analysis of perspectival structure. Employing that analysis and following the structural approach, I proposed three lines of empirical investigation that would selectively manipulate and measure vestibular processing and perspectival structure.

Day & Fitzpatrick (2005) quip that vestibular processes provide a "silent sense" (see also §2.2.1 of the target article). I suggested at the outset that (following the taxonomic approach) it might be surprisingly difficult to say with any precision why vestibular processing provides a sense of its very own. But even if it is true, that is, if the experiments described yield the expected results, they would show that vestibular processing is hardly silent. Indeed, each of the proposed lines of investigation would be a step towards a better understanding of how vestibular processes affect myriad forms of perspectival structure, all of which would further demonstrate the centrality of vestibular processing to our experiential life. In any case, my hope is that these remarks display the extent to which I have found Lenggenhager and Lopez's work to be not only inspirational, but also a rich and fruitful avenue for interdisciplinary research into the structural features of conscious experience.

⁹ The authors write that "deviations of 5°, 10°, and 15° lead to many erroneous self-attributions", found to be "decreasing in magnitude with increasing angular deviation" (Kannape et al. 2010, p. 1631). As broached above, one explanation of this pattern would be that deviations below 15° all fall (to a greater or lesser degree) within the phenomenal groove of the action specified by the task.

References

- Alsmith, A. (2012). What *reason*, consciousness in interaction: The role of the natural and social environment in shaping consciousness. In F. Paglieri (Ed.) (pp. 3-17) Amsterdam, NL: John Benjamins Press.
- Alsmith, A., Ferre, F., Haggard, P. & Longo, M. (unpublished). Isolating the origin(s) of perceptual perspective: The misalignment paradigm.
- Alsmith, A. & Longo, M. (2014). Where exactly am I? Self-location judgements distribute between head and torso. *Consciousness and Cognition*, 24, 70-74. [10.1016/j.concog.2013.12.005](https://doi.org/10.1016/j.concog.2013.12.005)
- Avillac, M., Denève, S., Olivier, E., Pouget, A. & Duhamel, J.-R. (2005). Reference frames for representing visual and tactile locations in parietal cortex. *Nature Neuroscience*, 8, 941-949. [10.1038/nn1480](https://doi.org/10.1038/nn1480)
- Blanke, O. & Metzinger, T. (2009). Full-body illusions and minimal phenomenal selfhood. *Trends in Cognitive Sciences*, 13 (1), 7-13. [10.1016/j.tics.2008.10.003](https://doi.org/10.1016/j.tics.2008.10.003)
- Boroditsky, L. (2000). Metaphoric structuring: Understanding time through spatial metaphors. *Cognition*, 75 (1), 1-28. [10.1016/S0010-0277\(99\)00073-6](https://doi.org/10.1016/S0010-0277(99)00073-6)
- Campbell, J. (1994). *Past, space, and self*. Cambridge, MA: MIT Press.
- Casasanto, D. & Boroditsky, L. (2008). Time in the mind: Using space to think about time. *Cognition*, 106 (2), 579-593. [10.1016/j.cognition.2007.03.004](https://doi.org/10.1016/j.cognition.2007.03.004)
- Day, B. L. & Fitzpatrick, R. C. (2005). The vestibular system. *Current Biology*, 15 (15), R583-R586. [10.1016/j.cub.2005.07.053](https://doi.org/10.1016/j.cub.2005.07.053)
- Ehrsson, H. (2007). The experimental induction of out-of-body experiences. *Science*, 317 (5841), 1048-1048. [10.1126/science.1142175](https://doi.org/10.1126/science.1142175)
- Fourneret, P. & Jeannerod, M. (1998). Limited conscious monitoring of motor performance in normal subjects. *Neuropsychologia*, 36 (11), 1133-1140. [10.1016/S0028-3932\(98\)00006-2](https://doi.org/10.1016/S0028-3932(98)00006-2)
- Gibson, J. J. (1950). *The perception of the visual world*. Boston, MA: Houghton Mifflin.
- Grubb, J. D. & Reed, C. L. (2002). Trunk orientation induces neglect-like lateral biases in covert attention. *Psychological Science*, 13 (6), 553-556. [10.1111/1467-9280.00497](https://doi.org/10.1111/1467-9280.00497)
- Horgan, T. E., Tienson, J. L. & Graham, G. (2003). The phenomenology of first-person agency. *Physicalism and mental causation* (pp. 323-340). Exeter, UK: Imprint Academic.
- Howard, I. P. (2012). *Perceiving in depth, volume 1: Basic mechanisms*. Oxford, UK: Oxford University Press.
- Ionta, S., Heydrich, L., Lenggenhager, B., Mouthon, M., Fornari, E., Chapuis, D. & Blanke, O. (2011). Multisensory mechanisms in temporo-parietal cortex support self-location and first-person perspective. *Neuron*, 70 (2), 363-374. [10.1016/j.neuron.2011.03.009](https://doi.org/10.1016/j.neuron.2011.03.009)
- Kannape, O. A., Schwabe, L., Tadi, T. & Blanke, O. (2010). The limits of agency in walking humans. *Neuropsychologia*, 48 (6), 1628-1636. [10.1016/j.neuropsychologia.2010.02.005](https://doi.org/10.1016/j.neuropsychologia.2010.02.005)
- Karnath, H. O., Schenkel, P. & Fischer, B. (1991). Trunk orientation as the determining factor of the 'contralateral' deficit in the neglect syndrome and as the physical anchor of the internal representation of body orientation in space. *Brain*, 114, 1997-2014. [10.1093/brain/114.4.1997](https://doi.org/10.1093/brain/114.4.1997)
- Knoblich, G. & Kircher, T. T. J. (2004). Deceiving oneself about being in control: Conscious detection of changes in visuomotor coupling. *Journal of Experimental Psychology: Human Perception and Performance*, 30 (4), 657-666. [10.1037/0096-1523.30.4.657](https://doi.org/10.1037/0096-1523.30.4.657)
- Lenggenhager, B., Tadi, T., Metzinger, T. & Blanke, O. (2007). Video ergo sum: Manipulating bodily self-consciousness. *Science*, 317 (5841), 1096-1099. [10.1126/science.1143439](https://doi.org/10.1126/science.1143439)
- Longo, M. & Alsmith, A. (2013). Where is the ego in egocentric representation? *Perception ECVF Abstract Supplement*, 42, 53-53.
- Mach, E. (1896/1959). *The analysis of sensation and the relation of the psychical to the physical* (C. M. Williams, Trans.). New York: Dover Publications.
- Macpherson, F. (2011a). *The senses: Classical and contemporary readings*. Oxford, UK: Oxford University Press.
- (2011b). Taxonomising the senses. *Philosophical Studies*, 153 (1), 123-142. [10.1007/s11098-010-9643-8](https://doi.org/10.1007/s11098-010-9643-8)
- Marcel, A. J. (2006). The sense of agency: Awareness and ownership of action. In J. Roessel & M. Eilan (Eds.) *Agency and self-awareness: Issues in philosophy and psychology* (pp. 48-93). Oxford, UK: Oxford University Press.
- Merleau-Ponty, M. (Ed.) (2002). *Phenomenology of perception*. London, UK: Routledge.
- Metzinger, T. (2003). *Being no one: The self-model theory of subjectivity*. Cambridge, MA: MIT Press.
- (2009). Self models. *Scholarpedia*, 2, 4174-4174.
- Mittelstaedt, H. (1992). Somatic versus vestibular gravity reception in man. *Annals of the New York Academy of Sciences*, 656 (1), 124-139. [10.1111/j.1749-6632.1992.tb25204.x](https://doi.org/10.1111/j.1749-6632.1992.tb25204.x)
- (1996). Somatic graviception. *Biological Psychology*, 42 (1-2), 53-74. [10.1016/0301-0511\(95\)05146-5](https://doi.org/10.1016/0301-0511(95)05146-5)

- Moll, H. & Meltzoff, A. N. (2011). Perspective taking and its foundation in joint attention. *Perception, causation and objectivity* (pp. 286-304). Oxford, UK: Oxford University Press.
- Peacocke, C. (1992). *A study of concepts*. Cambridge, MA: MIT Press.
- (1999). *Being known*. New York, NY: Oxford University Press.
- Perry, J. (1993). *The problem of the essential indexical and other essays*. Oxford, UK: Oxford University Press.
- Petkova, V. I. & Ehrsson, H. H. (2008). If I were you: Perceptual illusion of body swapping. *PLoS ONE*, 3 (12), e3832. [10.1371/journal.pone.0003832](https://doi.org/10.1371/journal.pone.0003832)
- Petkova, V. I., Khoshnevis, M. & Ehrsson, H. H. (2011a). The perspective matters! Multisensory integration in ego-centric reference frames determines full body ownership. *Frontiers in Psychology*, 2 (35), 1-7. [10.3389/fpsyg.2011.00035](https://doi.org/10.3389/fpsyg.2011.00035)
- Petkova, V. I., Björnsdotter, M., Gentile, G., Jonsson, T., Li, T.-Q. & Ehrsson, H. H. (2011b). From part-to whole-body ownership in the multisensory brain. *Current Biology*, 21 (13), 1118-1122. [10.1016/j.cub.2011.05.022](https://doi.org/10.1016/j.cub.2011.05.022)
- Pfeiffer, C., Lopez, C., Schmutz, V., Duenas, J. A., Martuzzi, R. & Blanke, O. (2013). Multisensory origin of the subjective first-person perspective: Visual, tactile, and vestibular mechanisms. *PLoS ONE*, 8 (4), e61751. [10.1371/journal.pone.0061751](https://doi.org/10.1371/journal.pone.0061751)
- Pfeiffer, C., Serino, A. & Blanke, O. (2014). The vestibular system: A spatial reference for bodily self-consciousness. *Frontiers in Integrative Neuroscience*, 8, 1-13. [10.3389/fnint.2014.00031](https://doi.org/10.3389/fnint.2014.00031)
- Serino, A., Alsmith, A., Costantini, M., Mandrigin, A., Tajadura-Jimenez, A. & Lopez, C. (2013). Bodily ownership and self-location: Components of bodily self-consciousness. *Consciousness and Cognition*, 22 (4), 1239-1252. [10.1016/j.concog.2013.08.013](https://doi.org/10.1016/j.concog.2013.08.013)
- Slachevsky, A., Pillon, B., Fournier, P., Pradat-diehl, P., Jeannerod, M. & Dubois, B. (2001). Preserved adjustment but impaired awareness in a sensory-motor conflict following prefrontal lesions. *Journal of Cognitive Neuroscience*, 13 (3), 332-340. [10.1162/08989290151137386](https://doi.org/10.1162/08989290151137386)
- Smith, A. J. T. (2010). Comment: Minimal conditions on the simplest form of self-consciousness. In T. Fuchs, H. Sattel & P. Henningsen (Eds.) *The embodied self: Dimensions, coherence, disorders* (pp. 35-41). Stuttgart, GER: Schattauer.
- Vaitl, D., Mittelstaedt, H., Saborowski, R., Stark, R. & Baisch, F. (2002). Shifts in blood volume alter the perception of posture: further evidence for somatic graviception. *International Journal of Psychophysiology*, 44 (1), 1-11. [10.1016/S0167-8760\(01\)00184-2](https://doi.org/10.1016/S0167-8760(01)00184-2)
- Vogeley, K. & Fink, G. R. (2003). Neural correlates of the first-person-perspective. *Trends in Cognitive Sciences*, 7 (1), 38-42. [10.1016/S1364-6613\(02\)00003-7](https://doi.org/10.1016/S1364-6613(02)00003-7)
- Zahavi, D. (2005). *Subjectivity and selfhood: Investigating the first-person perspective*. Cambridge, MA: MIT Press.

Vestibular Sense and Perspectival Experience

A Reply to Adrian Alsmith

Bigna Lenggenhager & Christophe Lopez

To answer Alsmith's questions about the existence of a vestibular sense, we outline in the first part of our reply why we believe the vestibular sense is a true "sixth sense". We argue that vestibular information constitutes distinct sensory events and that absolute coding of body orientation and motion in the gravity-centered space is the important unique feature of the vestibular system. In the last part of our reply, we extend Alsmith's experimental suggestions to investigate the vestibular contribution to various perspectival experiences.

Keywords

Absolute coding | Gravity | Otoliths | Perspective | Self-motion | Vestibular sense | Vestibular thresholds | Vestibular-evoked potentials

Authors

[Bigna Lenggenhager](#)

bigna.lenggenhager@usz.ch

University Hospital
Zurich, Switzerland

[Christophe Lopez](#)

christophe.lopez@univ-amu.fr

Aix Marseille Université
Marseille, France

Commentator

[Adrian Alsmith](#)

adrianjtalsmith@gmail.com

Københavns Universitet
Copenhagen, Denmark

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Is there a vestibular sense?

The first section of Alsmith's commentary ("Structural vs. taxonomic approaches to vestibular processes") raises an important question: *is there a vestibular sense?* The enduring lack of a clear answer to this seemingly simple question might stem from the old assumption that there are five and only five senses, all of

which giving rise to a distinct conscious sensation. The relatively late identification of the anatomical structures that code self-motion ([Wade 2003](#); [Lopez & Blanke 2014](#)) has probably further contributed to the neglect of the "vestibular sense" in philosophy and science. We comment below on two questions raised by

Alsmith concerning this debate: (1) *Are vestibular events sensory events?* and (2) *Are vestibular events of a specific kind, i.e., distinct from other sensations?*

(1) Are vestibular events sensory events? Several criteria have been proposed to determine whether an event is sensory or not (Macpherson 2011).¹ Following this type of approach, vestibular events can be described as sensory events because a sensory organ is dedicated to coding gravito-inertial forces and because there is a phenomenal experience associated with vestibular stimulation. Indeed, there are many situations during which passive own-body motions are characterized by distinct self-motion sensations. Imagine, for example, a situation in which we are sitting with eyes closed in the train and feel the departure, or when we are standing with eyes opened in a lift and experience vertical movement of the body. In such situations visual and somatosensory signals do not (or only weakly) contribute, but changes in vestibular signaling result in the conscious perception of self-motion, i.e., of “being translated forward” or “being elevated”.

Self-motion perception due to vestibular stimulation is also testable in the laboratory using motorized motion platforms (rotating chairs or translational platforms, see Palla & Lenggenhager 2014): participants are usually tested sitting on a chair, while non-vestibular sensory signals are largely excluded by having the participant’s body strapped to the chair and stabilized with cushions, by testing participants with eyes closed, by reducing auditory cues via white noise presented in headphones, and by testing participants with gloves and long sleeves (e.g., Grabherr et al. 2008; Hartmann et al. 2013; Lopez et al. 2013; Macaudo et al. 2014; Valko et al. 2012). Participants are able to accurately detect and report self-motion and its direction, which forms the basis for the measurement of *vestibular thresholds*, which are comparable to auditory or tactile thresholds. When accelerations are

applied above the threshold of the mechanoreceptors in the inner ear (e.g., above $0.6^\circ/\text{s}^2$ for rotations around the vertical axis), a motion sensation emerges in healthy participants, which in our opinion is the sensory event corresponding to the vestibular sensation “I was moved”. Such sensory events therefore constitute the basis of what has often been referred to as the “sixth sense” (Goldberg et al. 2012; Wade 2003; Berthoz 2000). Further compelling support comes from patients with dysfunctions of the peripheral vestibular apparatus like benign paroxysmal positional vertigo, vestibular neuritis, or Menière’s disease, who experience strong vestibular sensations in the form of vertigo (Brandt 1999).

We acknowledge, however, that in situations where we actively move the head with eyes opened in space, vestibular signals from self-motion do not give rise to such distinct “vestibular” sensation of self-motion. As explained in our target article, in conditions of active, self-generated head movements, vestibular signals are cancelled or strongly attenuated in the vestibular nuclei (Cullen 2011; Roy & Cullen 2004). This is probably why the vestibular sense has been termed a “silent sense” by some authors (Day & Fitzpatrick 2005).

(2) Are vestibular sensory events of a specific kind, i.e., distinct from other sensations? Vestibular sensations are sensations of own-body rotations, translations, and orientation (sensation of whole-body orientation with respect to the vertical) in space. Such sensations may in principle also emerge from the stimulation of other sensory systems, such as the visual, somatosensory and auditory systems. Impressively, illusory self-motion might be evoked by large optic flows, tactile stimulation under the feet, or displacement of auditory stimuli (Berthoz et al. 1975; Dichgans et al. 1972; Lackner & DiZio 2001, 2005; Våljamäe 2009). These findings resulted in Alsmith’s claim that “one may begin to seriously consider the possibility that vestibular processing does not constitute a form of sensory processing of its own kind” (this collection, p. 2). Yet if vestibular processing does

¹ For example, according to Macpherson, four main approaches to describe the senses can be distinguished: “the representational criterion,” “the phenomenal character criterion,” “the proximal stimulus criterion,” and “the sense-organ criterion” (2011).



Figure 1: A) *Crise de désinvolture* (2003) an artwork by Philippe Ramette. Copyrights: © 2015, ProLitteris, Zurich. All rights are reserved. Reproduction and any other use without permission - except for the individual and private use - is prohibited. B) Drawing of the “haunting sway”, a “gravity-defying” device that was originally developed in the US in the 1890s for amusement parks. The visitors had the impression that they were turning with the sway, while actually the room was turning around them.

not constitute a distinct form of sensory perception, to which type of sensory processing does it belong? Some authors have proposed that vestibular processing might relate to *proprioception* (since the vestibular system detects own body motions) or to *exteroception* (since it detects gravitational acceleration), but these propositions link vestibular processing to a function rather than a sensory modality. As recently pointed out by Macpherson (2011), “it is not even clear which sensory modality equilibration should be assimilated to, if indeed it should be assimilated to any” (p. 18).

Although vestibular, visual, and somesthetic signals may all support self-motion perception, this does not mean that the phenomenal experience of self-motion based on vestibular signals is similar to the experience based on visual signals. Actually, they may strongly differ in their content since, for example, the vestibular system is specialized in coding

high-frequency movements whereas the visual system is tuned to low-frequency movements (see also next paragraph).² And even at the neurophysiological level, vestibular signals interact very early with visual and somatosensory signals; yet this does not mean that these signals provide the exact same sensation of body motion and orientation. An analogy might be when we observe a person speaking: both auditory and visual signals from the speaker’s lip movements contribute to the experience of listening to a voice; nevertheless both signals provide clearly distinct sensations and experiences. We believe the same holds for vestibular processing. Vestibular sensations might be clearly distinct sensations, but in daily life they are often integrated with other senses, confounding a pure conscious sensation (Angelaki & Cullen 2008; Angelaki et al.

² We add that while visual, auditory, and somatosensory signals about self-motion can be suppressed, vestibular signals about body accelerations are *necessarily present*.

2009). Vestibular-only neurons are found in the vestibular nuclei, which are not influenced by visual signals or eye movements, suggesting that vestibular signals are not entirely fused with other sensory signals (Goldberg et al. 2012). Similarly, intracranial stimulations in epileptic patients have showed that pure vestibular sensations could be evoked during electrical stimulations of the superior temporal cortex and insula (Penfield 1957; Kahane et al. 2003; Mazzola et al. 2014).

2 A unique feature of the vestibular system: The representation of absolute self-motion and orientation

As mentioned by Alsmith, the vestibular system, unlike other sensory systems, does not code unique properties of sensory inputs such as loudness or hue. Yet, as already argued in the target article, the coding of *absolute* self-motion in space and self-orientation within gravity-related space is unique to the vestibular system. While relative (self-) motion and orientation can be detected by other sensory systems (e.g., vision and proprioception), gravity itself is not directly visible to these senses.³ Because vestibular organs contain gravito-inertial sensors, they provide a coding of body translations and rotations that is independent from external references (unlike visual, auditory, and somatosensory coding of whole-body motions). For this reason, vestibular organs code self-motion even when the eyes are closed, while we are jumping on a trampoline, or swimming in the sea.

With these properties the vestibular system, especially otolith signaling, also gives us the sensation of an “up” and a “down” by encoding gravitational acceleration. This process might be less accessible to consciousness in normal circumstances, as gravitational pull is constantly acting on vestibular mechanoreceptors. However, there is a large body of data showing that an “internal model of gravity” (predicting how objects move in the physical world according to Newton’s laws; McIntyre et al. 2001) which is strongly

based on otolith processing, shapes at a pre-conscious level several aspects of the visual perception of objects, body movements, and structure (e.g., Indovina et al. 2005; Lacquaniti et al. 2013; Lopez et al. 2009; Maffei et al. 2015; Yamamoto & Yamamoto 2006). A further illustration of the importance of the coding of body orientation in a gravity-centered space can be provided by the “tilted room illusion,” in which the furniture is aligned in a way that is incongruent with gravitational vertical (see figure 1A for an example by the French artist Philippe Ramette⁴), which has been used in a moving version as well in theme parks (the haunting swing, a “gravity-defying” ride, see figure 1B). Experiments conducted in this type of tilted environment have shown that the participant’s perception and posture are biased by tilted visual references, but not totally (Jenkin et al. 2003; Oman 2003). Merleau-Ponty has nicely noted the ambiguity of space-coding regarding the experience of *up* and *down*: “A direction can only exist for a subject who traces it out, and although a constituting mind eminently has the power to trace out all directions in space, in the present moment this mind has no direction and, consequently, it has no space, for it is lacking an actual starting point or an absolute here that could gradually give a direction to all determinations of space” (2012). It is interesting to note Merleau-Ponty’s claim that what is missing for the experience of up and down is an “absolute”. Merleau-Ponty also explains that “[w]e cannot, then, understand the experience of space through the consideration of the contents, nor through that of a pure activity of connecting, and we are confronted by that ‘*third spatiality*’ that we foreshadowed above, which is neither the spatiality of things in space, nor that of spatializing space [...] We need an ‘*absolute within the relative*’, a space that does not skate over appearances, that is anchored in them and depends upon them” (2012, p. 296–297; our italics). Although Merleau-Ponty did not mention the vestibular system when he described the necessity of a “third spatiality,” we now know that the otolithic sys-

³ Of course we can infer about (the direction of) gravity by the relative motion and specific properties of certain objects; however this process is much slower, less intuitive, and not always applicable.

⁴ To be precise, Ramette does not glue the furniture to the roof or wall, but rather “glues” himself to the wall. His position is thus tilted compared to gravity, not the furniture.

tem provides the “absolute within the relative” he mentions and allows the coding of absolute self-orientation in space (see also [Berthoz 2011](#) for a detailed account).

3 Vestibular system and perspectival experience—Experimental suggestions

In this last part we elaborate on the experimental suggestions provided by by Alsmith, proposed in order to investigate more fine-grained forms of perspectival perceptions and their interaction with vestibular processes. In the target article we used the term first-person perspective (mainly in the context of mental perspective taking and out-of-body experiences) to refer to an egocentric visuo-spatial perspective. Alsmith proposes a subdivision of this perspective into three forms of perspectival structures: “origin,” “egocentric frame of reference,” and “focal point of sensory flow (egomotion),” which might be differentially influenced by vestibular signals. While we do not necessarily agree on the importance and justification of these (and *exactly these*) components, we appreciate the experimental suggestions, on which we will briefly comment below.

3.1 Experiments I and II: Changing vestibular processes through change in perspective

A common approach to testing the influence of the vestibular system on high-level cognition is to alter vestibular information during a specific task—for example a perspective-taking task. This can be done either by applying galvanic ([Lenggenhager et al. 2008](#)) or caloric ([Falconer & Mast 2012](#)) vestibular stimulation, by natural vestibular stimulation ([Van Elk & Blanke 2014](#)), by exposing participants to microgravity ([Grabherr et al. 2007](#)), by changing the body orientation relative to gravity ([Arzy et al. 2006](#)), or by testing patients with vestibular dysfunction ([Grabherr et al. 2011](#)). What Alsmith describes⁵ in the first two experiments

mentioned in the commentary is the opposite approach, namely assessing vestibular processing during specific tasks, or bodily states, respectively.⁶ We believe that this is a potentially powerful way to better understand vestibular implication in fine-grained aspects of the bodily self and their interrelation—both in experimental work and research in patients with bodily-self disturbances (see e.g., [Brugger & Lenggenhager 2014](#) for a recent review). We would like, however, to point out a few important issues that should be considered.

Alsmith suggests that we measure time-locked vestibular-evoked potentials without stating more precisely what vestibular stimulation to use. However, this is crucial, since there are various ways to test vestibular processing, mostly by stimulating a specific part of the vestibular system (see e.g., [Palla & Lenggenhager 2014](#) for a recent review). One possibility (in the suggested experiment) could be to use sound-induced vestibular-evoked potentials. The advantage of these is that they can be recorded in a static condition, unlike other forms of vestibular stimulation (e.g., rotatory evoked cortical potentials; [Keck 1990](#)), which is important for the suggested full-body illusion paradigms. When designing experiments along these lines, it is indispensable to know what part of the vestibular system is stimulated by the used technique. Sound-induced cortical vestibular potentials, for example, represent cortical processing of otolith signals, mainly from the saccule, thus coding preferentially linear movements in the vertical plane (i.e., up and down movements in a standing position). If we rather expect a difference in coding the front-back movement, as proposed in Experiments 1 and 2, a vestibular stimulation of the utricle might be more appropriate (e.g., [Todd et al. 2014](#), using evoked-potentials by impulsive accelerations). Since testing all different aspects in all the proposed conditions is technically impossible, the specific vestibular stimulation should be carefully chosen based on the hypothesis. Alternatively,

⁵ This idea of measuring vestibular processes during situations of altered sense bodily self evolved in the framework of a grant entitled “Finding Perspective” awarded to Adrian Alsmith, Christophe Lopez and colleagues by the Volkswagen Foundation.

⁶ A similar approach has been used for other sensory processes such as the measure of body temperature during the rubber hand illusion ([Lenggenhager et al. 2014](#); [Moseley et al. 2008](#)) or the full-body illusion ([Macauda et al. 2014](#); [Salomon et al. 2013](#)).

more indirect measures could be used to test a vestibular implication, such as changes in posture or stability during various experimentally-induced alterations in the bodily self, e.g., via dynamic posturography using a moving platform, as it is commonly used in clinical settings (e.g., Ghulyan et al. 2005).

3.2 Experiment III: Egocentric perspective

In the third proposed experiment, Alsmith considers which (bodily) reference (e.g., eye, head or body centered) is taken as the egocentric reference frame. The fact that there are multiple bodily frames of reference has been nicely shown in a classical task where ambiguous letters (e.g., d/p) are written on the skin. They are typically perceived differently depending on the bodily location on which they are written (Sekiyama 1991); and interestingly the perspective can be modified by vestibular stimulation (Ferrè et al. 2014). Alsmith here suggests that there is a need to investigate the egocentric perspective both with implicit and explicit measures in a situation where body and head⁷ are misaligned, as previously done to test spatial cognition (Schindler 1997) and heading direction during passive motion (Ni et al. 2013). This is a very interesting suggestion; however from the experimental description it is not entirely clear how Alsmith thinks that the vestibular contribution should be investigated. Furthermore, his hypothesis only concerns the respective contribution of head and torso position, but not its vestibular contribution. He suggests that participants might receive galvanic vestibular stimulation or tendon vibration stimulation to investigate “the relative contribution of vestibular processes to egocentric perspective.” One way to test this could be to align the participant’s head and torso, but use tendon vibration or galvanic vestibular stimulation in order to induce an illusory tilt or turn the participant’s head, thus inducing an illusory misalignment of the head and body. By doing the suggested task in such a condition, vestibular or proprioceptive contribution could be isolated.

While this is theoretically very interesting, there might be practical difficulties: vestibular and proprioceptive illusions are usually susceptible to huge individual differences, and inducing illusory shift of $\pm 15\%$ could be difficult. Furthermore, in the proposed experiment that misaligns body and head around the yaw axis, gravitational cues do not differ between the position of the torso and the head in the misaligned condition. Adapting the experiment to a lying-down position,⁸ where body and head would be at different angles with respect to gravity, could help investigating the otolithic influence on perspective.

4 Conclusion

In response to Alsmith’s inspiring theoretical suggestions, we have argued that there is a true vestibular sense, with distinct and important properties. We believe and agree with Alsmith that better understanding its contribution to various aspects of experiential life is crucial and that this might also facilitate taxonomic and structural approaches. Alsmith’s response exemplifies, in our view, the mutual benefit of an interdisciplinary dialogue, as his thorough analysis of current experimental data, paired with new theoretical considerations, leads to concrete experimental suggestions, which might reshape theoretical considerations depending on the potential results. In our reply we have pointed out some possible methodological difficulties, some possible ways to overcome these, and some new directions such experimental work could take. In particular, we are optimistic that analyzing vestibular processing in the brain using electrophysiological approaches will provide in the near future important new data about the vestibular contribution to the sense of self. We hope that our reply will help foster interdisciplinary collaborations that further investigate the role of the vestibular system in shaping our mind.

⁷ Additionally, eye-position could be manipulated.

⁸ Or generally test various body orientations (e.g., as in Lopez et al. 2009).

References

- Alsmith, A. (2015). Perspectival Structure and Vestibular Processing. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Angelaki, D. E. & Cullen, K. E. (2008). Vestibular system: The many facets of a multimodal sense. *Annual Review of Neuroscience*, 31, 125-150. [10.1146/annurev.neuro.31.060407.125555](https://doi.org/10.1146/annurev.neuro.31.060407.125555)
- Angelaki, D. E., Klier, E. M. & Snyder, L. H. (2009). A vestibular sensation: Probabilistic approaches to spatial perception. *Neuron*, 64, 448-461. [10.1016/j.neuron.2009.11.010](https://doi.org/10.1016/j.neuron.2009.11.010)
- Arzy, S., Thut, G., Mohr, C., Michel, C. M. & Blanke, O. (2006). Neural basis of embodiment: Distinct contributions of temporoparietal junction and extrastriate body area. *Journal of Neuroscience*, 26 (31), 8074-8081. [10.1523/JNEUROSCI.0745-06.2006](https://doi.org/10.1523/JNEUROSCI.0745-06.2006)
- Berthoz, A. (2000). *The brain's sense of movement*. Cambridge, MA: Harvard University Press.
- (2011). La conscience du corps. In A. Berthoz & B. Andrieu (Eds.) *Le corps en acte: Centenaire Maurice Merleau Ponty* (pp. 9-22). Nancy, F: Presses Universitaires de Nancy.
- Berthoz, A., Pavard, B. & Young, L. R. (1975). Perception of linear horizontal self-motion induced by peripheral vision (linearvection) basic characteristics and visual-vestibular interactions. *Experimental Brain Research*, 23 (5), 471-489. [10.1007/BF00234916](https://doi.org/10.1007/BF00234916)
- Brandt, T. (1999). *Vertigo. Its multisensory syndromes*. London, UK: Springer.
- Brugger, P. & Lenggenhager, B. (2014). The bodily self and its disorders: Neurological, psychological and social aspects. *Current Opinion in Neurology*, 27 (6), 644-652. [10.1097/WCO.0000000000000151](https://doi.org/10.1097/WCO.0000000000000151)
- Cullen, K. E. (2011). The neural encoding of self-motion. *Current Opinion in Neurobiology*, 21 (4), 587-595. [10.1016/j.conb.2011.05.022](https://doi.org/10.1016/j.conb.2011.05.022)
- Day, B. L. & Fitzpatrick, R. C. (2005). The vestibular system. *Current Biology*, 15 (15), R583-R586. [10.1016/j.cub.2005.07.053](https://doi.org/10.1016/j.cub.2005.07.053)
- Dichgans, J., Held, R., Young, L. R. & Brandt, T. (1972). Moving visual scenes influence the apparent direction of gravity. *Science*, 178, 1217-1219.
- Falconer, C. J. & Mast, F. W. (2012). Balancing the mind: Vestibular induced facilitation of egocentric mental transformations. *Experimental Psychology*, 59 (6), 332-339. [10.1027/1618-3169/a000161](https://doi.org/10.1027/1618-3169/a000161)
- Ferrè, E. R., Lopez, C. & Haggard, P. (2014). Anchoring the self to the body: Vestibular contribution to the sense of self. *Psychological Science*, 25 (11), 2106-2108. [10.1177/0956797614547917](https://doi.org/10.1177/0956797614547917)
- Ghulyan, V., Paolino, M., Lopez, C., Dumitrescu, M. & Lacour, M. (2005). A new translational platform for evaluating aging or pathology-related postural disorders. *Acta Oto-Laryngologica*, 125 (6), 607-617. [10.1080/00016480510026908](https://doi.org/10.1080/00016480510026908)
- Goldberg, J. M., Wilson, V. J., Cullen, K. E., Angelaki, D. E., Broussard, D. M., Büttner-Ennever, J. A. & Minor, L. B. (2012). *The vestibular system. A sixth sense*. New York, NY: Oxford University Press.
- Grabherr, L., Karmali, F., Bach, S., Indermaur, K., Metzler, S. & Mast, F. W. (2007). Mental own-body and body-part transformations in microgravity. *Journal of Vestibular Research*, 17 (5-6), 279-287.
- Grabherr, L., Nicoucar, K., Mast, F. W. & Merfeld, D. M. (2008). Vestibular thresholds for yaw rotation about an earth-vertical axis as a function of frequency. *Experimental Brain Research*, 186 (4), 677-681. [10.1007/s00221-008-1350-8](https://doi.org/10.1007/s00221-008-1350-8)
- Grabherr, L., Cuffel, C., Guyot, J.-P. & Mast, F. W. (2011). Mental transformation abilities in patients with unilateral and bilateral vestibular loss. *Experimental Brain Research*, 209 (2), 205-214. [10.1007/s00221-011-2535-0](https://doi.org/10.1007/s00221-011-2535-0)
- Hartmann, M., Furrer, S., Herzog, M. H., Merfeld, D. M. & Mast, F. W. (2013). Self-motion perception training: Thresholds improve in the light but not in the dark. *Experimental Brain Research*, 226 (2), 231-240. [10.1007/s00221-013-3428-1](https://doi.org/10.1007/s00221-013-3428-1)
- Indovina, I., Maffei, V., Bosco, G., Zago, M., Macaluso, E. & Lacquaniti, F. (2005). Representation of visual gravitational motion in the human vestibular cortex. *Science*, 308 (5720), 416-419. [10.1126/science.1107961](https://doi.org/10.1126/science.1107961)
- Jenkin, H. L., Dyde, R. T., Jenkin, M. R., Howard, I. P. & Harris, L. R. (2003). Relative role of visual and non-visual cues in determining the direction of “up”: Experiments in the York tilted room facility. *Journal of Vestibular Research*, 13 (4-6), 287-293. [10.1016/j.actaastro.2005.01.030](https://doi.org/10.1016/j.actaastro.2005.01.030)
- Kahane, P., Hoffmann, D., Minotti, L. & Berthoz, A. (2003). Reappraisal of the human vestibular cortex by cortical electrical stimulation study. *Annals of Neurology*, 54 (5), 615-624.
- Keck, W. (1990). Rotatory evoked cortical potentials in normal subjects and patients with unilateral and bilateral vestibular loss. *European Archives of Oto-Rhino-Laryngology*, 247 (4), 222-225. [10.1007/BF00178989](https://doi.org/10.1007/BF00178989)
- Lackner, J. R. & DiZio, P. (2001). Somatosensory and proprioceptive contributions to body orientation, sensory localization, and self-calibration. In R. J. Nelson (Ed.) *The somatosensory system: Deciphering the brain's own body image* (pp. 121-140). Boca Raton, FL: CRC Press.

- (2005). Vestibular, proprioceptive, and haptic contributions to spatial orientation. *Annual Review of Psychology*, 56, 115-147. [10.1146/annurev.psych.55.090902.142023](https://doi.org/10.1146/annurev.psych.55.090902.142023)
- Lacquaniti, F., Bosco, G., Indovina, I., La Scaleia, B., Maffei, V., Moscatelli, A. & Zago, M. (2013). Visual gravitational motion and the vestibular system in humans. *Frontiers in Integrative Neuroscience*, 7 (101). [10.3389/fnint.2013.00101](https://doi.org/10.3389/fnint.2013.00101)
- Lenggenhager, B., Lopez, C. & Blanke, O. (2008). Influence of galvanic vestibular stimulation on egocentric and object-based mental transformations. *Experimental Brain Research*, 184 (2), 211-221. [10.1007/s00221-007-1095-9](https://doi.org/10.1007/s00221-007-1095-9)
- Lenggenhager, B., Hilti, L., Palla, A., Macaudo, G. & Brugger, P. (2014). Vestibular stimulation does not diminish the desire for amputation. *Cortex*, 54, 210-212. [10.1016/j.cortex.2014.02.004](https://doi.org/10.1016/j.cortex.2014.02.004)
- Lopez, C. & Blanke, O. (2014). Nobel Prize centenary: Robert Bárány and the vestibular system. *Current Biology*, 24 (21), R1026-R1028. [10.1016/j.cub.2014.09.067](https://doi.org/10.1016/j.cub.2014.09.067)
- Lopez, C., Bachofner, C., Mercier, M. & Blanke, O. (2009). Gravity and observer - body orientation influence the visual perception of human body postures. *Journal of Vision*, 9 (5), 1-14. [10.1167/9.5.1](https://doi.org/10.1167/9.5.1)
- Lopez, C., Falconer, C. J. & Mast, F. W. (2013). Being moved by the self and others: influence of empathy on self-motion perception. *PloS One*, 8 (1), e48293-e48293. [10.1371/journal.pone.0048293](https://doi.org/10.1371/journal.pone.0048293)
- Macaudo, G., Bertolini, G., Palla, A., Straumann, D., Brugger, P. & Lenggenhager, B. (2014). Binding body and self in visuo-vestibular conflicts. *European Journal of Neuroscience*, 13, 556-571. [10.1038/nrn3292](https://doi.org/10.1038/nrn3292)
- Macpherson, F. (2011). Individuating the senses. In F. Macpherson (Ed.) *The senses: Classic and contemporary philosophical perspectives* (pp. 3-43). Oxford, UK: Oxford University Press.
- Maffei, V., Indovina, I., Macaluso, E., Ivanenko, Y. P., A. Orban, G. & Lacquaniti, F. (2015). Visual gravity cues in the interpretation of biological movements: Neural correlates in humans. *NeuroImage*, 104, 221-230. [10.1016/j.neuroimage.2014.10.006](https://doi.org/10.1016/j.neuroimage.2014.10.006)
- Mazzola, L., Lopez, C., Faillenot, I., Chouchou, F., Mauguère, F. & Isnard, J. (2014). Vestibular responses to direct stimulation of the human insular cortex. *Annals of Neurology*, 76 (4), 609-619. [10.1002/ana.24252](https://doi.org/10.1002/ana.24252)
- McIntyre, J., Zago, M., Berthoz, A. & Lacquaniti, F. (2001). Does the brain model Newton's laws? *Nature Neuroscience*, 4 (7), 693-694. [10.1038/89477](https://doi.org/10.1038/89477)
- Merleau-Ponty, M. (2012). *Phenomenology of perception*. London, UK: Routledge.
- Moseley, G. L., Olthof, N., Venema, A., Don, S., Wijers, M., Gallace, A. & Spence, C. (2008). Psychologically induced cooling of a specific body part caused by the illusory ownership of an artificial counterpart. *Proceedings of the National Academy of Sciences, USA*, 105, 13169-13173. [10.1073/pnas.0803768105](https://doi.org/10.1073/pnas.0803768105)
- Ni, J., Tatalovic, M., Straumann, D. & Olasagasti, I. (2013). Gaze direction affects linear self-motion heading discrimination in humans. *European Journal of Neuroscience*, 38 (8), 3248-3260. [10.1111/ejn.12324](https://doi.org/10.1111/ejn.12324)
- Oman, C. M. (2003). Human visual orientation in weightlessness. *Levels of Perception* (pp. 375-398). New York, NY: Springer.
- Palla, A. & Lenggenhager, B. (2014). Ways to investigate vestibular contributions to cognitive processes. *Frontiers in Integrative Neuroscience*, 8 (40). [10.3389/fnint.2014.00040](https://doi.org/10.3389/fnint.2014.00040)
- Penfield, W. (1957). Vestibular sensation and the cerebral cortex. *Annals of Otology, Rhinology & Laryngology*, 66, 691-698.
- Roy, J. E. & Cullen, K. E. (2004). Dissociating self-generated from passively applied head motion: Neural mechanisms in the vestibular nuclei. *Journal of Neuroscience*, 24, 2102-2111. [10.1523/JNEUROSCI.3988-03.2004](https://doi.org/10.1523/JNEUROSCI.3988-03.2004) 24/9/2102
- Salomon, R., Lim, M., Pfeiffer, C., Gassert, R. & Blanke, O. (2013). Full body illusion is associated with widespread skin temperature reduction. *Frontiers in Behavioral Neuroscience*, 7 (65). [10.3389/fnbeh.2013.00065](https://doi.org/10.3389/fnbeh.2013.00065)
- Schindler, G. K. (1997). Head and trunk orientation modulate visual neglect. *Neuroreport*, 8 (12), 2681-2685. [10.1097/00001756-199708180-00009](https://doi.org/10.1097/00001756-199708180-00009)
- Sekiyama, K. (1991). Importance of head axes in perception of cutaneous patterns drawn on vertical body surfaces. *Perception and Psychophysics*, 49 (5), 481-492. [10.3758/BF03212182](https://doi.org/10.3758/BF03212182)
- Todd, N. P. M., McLean, A., Paillard, A., Kluk, K. & Colebatch, J. G. (2014). Vestibular evoked potentials (VsEPs) of cortical origin produced by impulsive acceleration applied at the nasion. *Experimental Brain Research*, 232 (12), 3771-3784. [10.1007/s00221-014-4067-x](https://doi.org/10.1007/s00221-014-4067-x)
- Valko, Y., Lewis, R. F., Priesol, A. J. & Merfeld, D. M. (2012). Vestibular labyrinth contributions to human whole-body motion discrimination. *Journal of Neuroscience*, 32 (39), 13537-13542. [10.1523/JNEUROSCI.2157-12.2012](https://doi.org/10.1523/JNEUROSCI.2157-12.2012)
- Van Elk, M. & Blanke, O. (2014). Imagined own-body transformations during passive self-motion. *Psychological Research*, 78 (1), 18-27. [10.1007/s00426-013-0486-8](https://doi.org/10.1007/s00426-013-0486-8)

- Väljamäe, A. (2009). Auditorily-induced illusory self-motion: A review. *Brain Research Reviews*, 61 (2), 240-255. [10.1016/j.brainresrev.2009.07.001](https://doi.org/10.1016/j.brainresrev.2009.07.001)
- Wade, N. J. (2003). The search for a sixth sense: The cases for vestibular, muscle, and temperature senses. *Journal of the History of the Neurosciences*, 12 (2), 175-202. [10.1076/jhin.12.2.175.15539](https://doi.org/10.1076/jhin.12.2.175.15539)
- Yamamoto, S. & Yamamoto, M. (2006). Effects of the gravitational vertical on the visual perception of reversible figures. *Neuroscience Research*, 55 (2), 218-221. [10.1016/j.neures.2006.02.014](https://doi.org/10.1016/j.neures.2006.02.014)

Self-as-Subject and Experiential Ownership

Caleb Liang

In what follows, I investigate the distinction between the sense of self-as-object and the sense of self-as-subject, and propose an account that is different from Shoemaker's immunity principle. I suggest that this distinction can be elucidated by examining two types of self-experience: the sense of *body ownership* and the sense of *experiential ownership*. The former concerns self-as-object: whether a body part or a full body belongs to me. The latter concerns self-as-subject: whether I represent myself as the unique subject of experience. A key point is that misrepresentation can occur not only in the sense of body ownership but also in the sense of experiential ownership. Then I examine the most relevant neuroscientific accounts of the sense of self-as-subject, including Damasio's account of the core-self, Panksepp's affective neuroscience, neural synchrony, and the sub-cortical-cortical midline structures. I argue that none of these successfully explains the neural basis of the sense of self-as-subject. In order to make progress, I suggest, the first step is to look for and then to study the various conditions in which one can pursue the "Wittgenstein Question".

Keywords

Body ownership | Core-self | Experiential ownership | Immunity principle | Neural synchrony | Self-as-object | Self-as-subject

Author

Caleb Liang

yiliang@ntu.edu.tw

National Taiwan University
Taipei, Taiwan

Commentators

Oliver Haug

ruehlo1@students.uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Marius F. Jung

mjung02@students.uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

This paper investigates a central form of self-consciousness from an interdisciplinary perspective: *the sense of self-as-subject*.¹ How philosophers understand this form of consciousness has been influenced by two ideas. One is Wittgenstein's distinction between "I"-as-object and

"I"-as-subject. In the *Blue Book* (1958), he says that: "there is no question of recognizing a person when I say I have toothache. To ask 'are you sure it is *you* who have pains?' would be nonsensical". The other is Shoemaker's immunity principle. Developing Wittgenstein's distinction, Shoemaker (1968) argues that we are "immune to error through misidentification relative to the first-person pronouns (IEM)". Many consider IEM to be solely addressing semantic

¹ Here I will focus on the minimal sense of self-as-subject, which means that the sense of self-as-subject does not require exercising conceptual capacities and can be transient. It is contrasted with the "narrative self" or "autobiographical self", which involves episodic memory and persists through time (Gallagher 2000).

or conceptual issues. But for philosophers of mind, it decisively sets apart two types of self-consciousness. When one is conscious of oneself-as-object, error is always possible; however, when one is conscious of oneself-as-subject, a particular sort of mistake about *who* the subject is becomes impossible.

The first goal of this paper is to propose an alternative explication of the sense of self-as-object and the sense of self-as-subject. I aim to provide an account that is both phenomenologically precise and empirically useful. The distinction, I will suggest, can be better understood as two types of self-experience: a sense of body ownership and a sense of experiential ownership. I will argue that sometimes it makes perfect sense to ask a subject “are you sure it is *you* who feels pain?” For brevity, I will call this type of question the “Wittgenstein Question”. I will also argue that IEM, or at least some versions of it, faces counterexamples from empirical research. The second goal of this paper is to examine empirical accounts related to the sense of self-as-subject. There are currently many neuroscience programs devoted to self-consciousness, and recently some researchers claim to have explained the neural mechanisms of the sense of self-as-subject. Investigating these programs will reveal how philosophy can contribute to neuroscience in understanding this target phenomenon.

I discuss the sense of body ownership in section 2, and explain how it helps to clarify the sense of self-as-object. Section 3 introduces the notion of experiential ownership. I use this notion to specify what it is like to experience the self-as-subject. A crucial claim is that *being* the subject of an experience does not imply experiencing oneself *as* the subject of experience. If this is correct, at least some forms of IEM fail. Consequently, if we want to talk about a sense of self-as-subject we need more empirical studies. Section 4 examines Damasio’s account of the core-self and Panksepp’s affective neuroscience. Both claim to explain the neural basis of the sense of self-as-subject, but I argue that they only address the sense of self-as-object. In section 5, I criticize two proposals that some neuroscientists use for explaining the sense of

self-as-subject: neural synchrony and subcortical-cortical midline structures (SCMS). The overall positive lesson we can take from these accounts will be presented in the final section.

2 Body ownership and self-as-object

The sense of body ownership concerns whether a body part or a whole body is experienced as belonging to me. For example, I am now typing this paper with two hands, and I have a sense that the two hands are mine. To clarify this concept of self-experience, three distinctions will be very useful. One is between the *fact* of body ownership and the *sense* of body ownership (Dokic 2003; de Vignemont 2011). The former is a biological fact about the anatomical structures of one’s body. The latter is a conscious experience of the fact of body ownership. As the syndrome of somatoparaphrenia indicates, these two aspects are dissociable. A prominent feature of somatoparaphrenia is that patients deny that parts of their body, e.g., a hand, belongs to them (Vallar & Ronchi 2009). Their sense of body ownership fails to match up with the facts—namely, that that the hand is theirs.

In healthy subjects, the sense of body ownership can also be mistaken. In the rubber hand illusion (RHI), participants experience a fake hand as belonging to them. The set-up is simple: The subject’s own hand is blocked from view. The subject sees a rubber hand in front of her, clearly distinct from her own real hand. The experimenter uses paint brushes to touch the real hand and the rubber hand either synchronously or asynchronously (Botvinick & Cohen 1998; Tsakiris & Haggard 2005). In the synchronous condition, many subjects report that they feel as though they are being touched on the rubber hand rather than on their real hand. More interestingly, many subjects feel as if the rubber hand were their own hand.²

Another form of misrepresentation involves the full body—an illusion that induces some interesting aspects of out-of-body experience

² Proprioceptive drift is another aspect frequently associated with RHI: many subjects judge (by proprioception) their real hand as being located closer to the rubber hand, rather than as where it really is. But Rohde et al. (2011) have recently shown that this aspect can be dissociated from the feeling of the rubber hand as one’s own.

(OBE) (Lenggenhager et al. 2007).³ In experiments of this type, the subject wears a three-dimensional head-mounted display (HMD), and a stereo camera stands two meters behind her. The scenes registered by the camera are transmitted to the HMD such that the subject sees the back of his virtual body in front of her. Then the subject's back is stroked either synchronously or asynchronously with the virtual body. In the synchronous condition, many subjects feel as if the virtual body were their own.⁴

The second distinction is between the first-personal sense and third-personal sense of body ownership. In daily experience, the sense of body ownership is often first-personal as well as pre-reflective (Legrand 2007, 2010). That is, by proprioception and somatosensation, I can experience the body as mine *from the inside* without watching it or reflecting upon it (de Vignemont 2012). Consider simple activities such as walking. When I talk to someone while walking, my attention can be fully absorbed in the conversation. In this case, I don't pay any attention to my leg movements. Still, due to the firing patterns of muscle spindles in my legs, I implicitly experience that my legs take turns entering into the stance phase (touching the ground) and the swing phase (leaving the ground) to move my body forward. In contrast, the sense of body ownership can sometimes be third-personal and reflective. When looking at a

monitor in an airport showing the image of my body, I may wonder whether the body that I see is mine. In this case, instead of experiencing it *from the inside*, I consider my body from the third-person point of view. That is, the body is treated as the object of visual experience, attention, or reflection.⁵ In the rest of this paper, I will use "the sense of body ownership" to indicate the first-personal sense of the term.⁶

These two distinctions have been suggested before. But now I want to propose a third distinction to help elucidate what we mean when we talk about the sense of self-as-object. This third distinction refers to the difference between a sense of body ownership and a sense of self *as a physical body*.⁷ The former relates to questions like "Is this my hand?" and "Is that body mine?", whereas the latter concerns issues such as "What am I?" and "Am I a physical object?" This distinction marks two notions of bodily self-consciousness: experiencing a body part or a full body as one's *own*, on the one hand, and being conscious of oneself *as a physical body* on the other. Conceptually, the sense of *having* a body and the sense of *being* a body are different notions.⁸ However, they are closely related *experientially*. I suggest that experien-

³ Cf. Ehrsson (2007) for a different OBE experiment.

⁴ The relationship between body-part and whole-body representations for body ownership is a controversial issue. Clearly they are not the same. The issue is: are they fundamentally different? Or is the difference only a matter of degree? As an anonymous reviewer points out, during the rubber-hand illusion, one's self-location and global body ownership are unaffected. However, during full-body illusions these aspects are affected and misrepresented because they concern the whole-body. Some researchers might therefore think that there exist some fundamental differences between body-part and whole-body representations for body ownership. One can also reasonably hypothesize that the neural mechanisms that are responsible for hand ownership do not need to involve brain regions that process leg or trunk representations. However, in my opinion more interdisciplinary studies would be required to really solve this issue. My current position is that, regarding the sense of body ownership, the difference between body-part and whole-body representations is a matter of degree. First, conceptually speaking, there doesn't seem to be a sharp distinction between body-part and whole-body representations. Second, if we consider the experimental set-ups of the rubber hand illusion and of the full-body illusions (either Lenggenhager's version or Ehrsson's), the differences between them seem to be a matter of degree as well. Of course, these are not arguments yet. I have recently designed a set of experiments precisely to deal with this issue, and I hope to be able to say something about it soon.

⁵ Are there borderline cases between the first-personal and the third-personal experiences of one's own body? I think so. For example, to use the above example again, if one of my legs suddenly hurts a little bit, I may be able to continue my conversation without disruption, but I have to pay attention to proprioception in order to walk normally. In this case, I submit, the distinction between the first-personal and the third-personal senses of body ownership is not sharp. However, this will not affect my proposal below regarding the relationship between the sense of full-body ownership and the sense of self-as-object.

⁶ Both the first-personal and the third-personal senses of body ownership are involved in RHI and OBE. On the one hand, the fake hand or the virtual body that the subject sees is the object of visual awareness, which is experienced as standing apart from their visual perspective. In addition, by filling in the questionnaires after the experiment, the subject makes explicit judgments about body ownership. This is the third-personal sense of body ownership. On the other hand, during the experiment, the synchronous touch and proprioception causes the subject to feel as if "it is my body that is being touched". This is the first-personal sense of body ownership, which can be indirectly measured by skin conductance response (SCR). In RHI and OBE, both the third-personal and the first-personal senses of body ownership are prone to misrepresentation.

⁷ Here, "physical body" is broadly construed such that it can refer not only to a physical object but also to a biological organism or a flesh-and-blood person.

⁸ A Cartesian dualist might say that, although I experience a particular body as mine, I fundamentally conceive of myself *as* a thinking being rather than as a physical body. For the purpose of this paper, we can set Cartesianism aside.

cing ownership of a full body provides a sense of self as a physical body. When I engage in daily activities, there is not only a sense that this body is mine but also a sense that I *am* a physical body. Consider ordinary experiences like eating, running, bleeding, standing behind a desk, etc. These experiences involve a sense of body ownership, i.e. what it is like to *have* a body. But I also experience what it is like to *be* something that is eating, running, bleeding, etc. That is, I have a sense about *what* I am, or a sense of myself as a physical body that is doing these things.

I suggest that the sense of full-body ownership helps us to understand the sense of self-as-physical-body.⁹ The sense of self-as-physical-body, in turn, helps us to specify what it means to be conscious of the self-as-object.¹⁰ When I experience these hands as mine, there is a sense in which I am implicitly aware of myself *as* a physical body such that these two hands *are parts of me*. The proposal here is that I am conscious of myself-*as-object* when I am conscious of myself as a physical body. This holds not only in cases where I take myself as an object of vision or attention, such as seeing myself in a mirror. It holds even when I experience myself as a body from the first person perspective.¹¹

⁹ The idea is that we know how to conduct empirical research in order to study the sense of full-body ownership which, as Blanke and Metzinger suggest, is connected with the following features: (i) the global sense of identification with a physical body as a whole (self-identification); (ii) the sense of being situated in a specific place (self-location); and (iii) the sense of possessing “a point of projection functioning as its origin in sensory and mental processing (weak 1PP)” (2009, pp. 7–8). Together, these features characterize what Blanke and Metzinger call minimal phenomenal selfhood (MPS), defined as “the conscious experience of being a what” (2009, p. 7). It is my view that these three features articulate what it is like to be a self *as a physical body*. In this regard, the sense of full-body ownership helps us to understand the sense of self-as-physical-body. Also, thanks to the recent findings of the RHI and the OBE experiments, we have now better ideas regarding how misrepresentation may occur in the sense of body ownership. This, in turn, suggests that the sense of self-as-physical-body can involve misrepresentation as well.

¹⁰ In my account, “the sense of self-as-physical-body” serves as a conceptual bridge between “the sense of full-body ownership” and the “sense of self-as-object”. Experientially, the sense of full-body ownership and the sense of self-as-physical-body are closely related. I deliberately leave open whether these two notions denote the same or different experiences. I think more interdisciplinary work will be required to fix this issue.

¹¹ My proposal here is very different from what might be called the Pre-reflective Account of self-consciousness (Legrand 2006, 2007, 2010, 2011; Gallagher 2005; Zahavi 2005). According to this account, self and body are constitutively tied together, and body can provide a sense of self-as-subject, i.e., one can experience one’s body-*as-sub-*

ject. Let me draw some remarks made by Wittgenstein to support this proposal. Consider his examples of “I”-as-object: “My arm is broken”, “I have grown six inches”, “I have a bump on my forehead” (1958, p. 67). These examples clearly refer to the speaker’s body. This fits my suggestion that consciousness of self-as-object can be understood as consciousness of self-as-physical-body—I have the sense that I am a body that has a broken arm or that has grown six inches. Now consider his examples of “I”-as-subject: “I see so-and-so”, “I try to lift my arm”, “I have toothache” (1958, pp. 66–67). As indicated by his own italicization, the use of “I”-as-subject is about *who* the perceiver, agent, or the subject is. But notice that these examples refer to the speaker’s body *as well*. What does this tell us? My interpretation is that it implies that the idea of *who the subject is* should not be regarded as the same as the idea of *what* does the perceiving, lifting, or undergoes toothache. The sense of self-as-subject is not equivalent to the sense of self-as-physical-body.

Towards the end of *The Blue Book*, Wittgenstein makes two important remarks. First, “we can perfectly well adopt the expression “this body feels pain”, and we shall then, just as usual, tell it to go to the doctor, to lie down, and even to remember that when the last time it had pains they were over in a day” (1958, p. 73).¹² His point is that we should not construe the thing that suffers pain as a Cartesian immaterial ego. The notion of body in the expression

ject. Pre-reflectively experiencing the self as a physical body would correspond to the sense of body-as-subject rather than as-object. The difference between my view and this account centers on whether the notion of object in “self-as-object” is construed as a physical body or as an “intentional object of consciousness”. I contend that the sense of self-as-subject is different from the sense of body-as-subject. Experiencing the self as the subject of experiences is not the same as experiencing the self as a perceiving or acting body. I address these issues in another paper.

¹² Just before this, Wittgenstein says: “Let us now ask: ‘Can a human body have pain?’ One is inclined to say: ‘How can the body have pain? The body in itself is something dead; a body isn’t conscious!’ And here again it is as though we looked into the nature of pain and saw that it lies in its nature that a material object can’t have it. And it is as though we saw that what has pain must be an entity of a different nature from that of a material object; that, in fact, it must be of a mental nature. But to say that the ego is mental is like saying that the number 3 is of a mental or an immaterial nature, when we recognize that the numeral ‘3’ isn’t used as a sign for a physical object” (1958, p.73).

‘this body feels pain’ can perfectly well refer to a physical object, i.e. to a person or to a biological organism that can consciously feel pain. Wittgenstein states this point from the third-person perspective. But there is no reason why this point cannot be formulated from the first-person perspective. That is, by “this body” I can refer to myself. As I suggested above, I can experience my body *from the inside*. Someone else can tell me to go to the doctor or to lie down, etc. In this case, I can be aware of myself as *having* a body that is in pain (a sense of body ownership), and I can have a sense of myself *as* a body that is in pain (the sense of self-as-physical-body).

This brings us to Wittgenstein’s second remark: “The kernel of our proposition that that which has pains or sees or thinks is of a mental nature is only that the word ‘I’ in ‘I have pains’ does not denote a particular body, for we can’t substitute for ‘I’ a description of a body” (1958, pp. 73–74). My interpretation of this remark is that, even when it is my body that is in pain, there remains a difference between saying “I have pains” and saying “this particular body feels pain”. When Wittgenstein says that “the word ‘I’ in ‘I have pains’ does not denote a particular body”, this remark can apply to the speaker’s body considered *from the first-person perspective*. The reason why we can’t substitute for “I” a description of a body is *not* because my body has to be described from the third-person point of view or that it has to be treated as an intentional object of consciousness. Rather, the reason we can’t substitute for “I” a description of a body is that the “I” in “I have pains” captures the sense of *who* feels pains, while “a particular body” captures the sense of *what* feels pains. This difference, then, marks two different types of self-consciousness. In the former case, I am conscious of myself as the subject of pain experience. In the latter case, I am conscious of myself *as the body* that feels pain. I do not mean that this is the only possible interpretation of Wittgenstein’s remarks. My claim is that it is a plausible interpretation, according to which the sense of self as subject of experience is

distinct from the sense of self as a physical body, even when the body is characterized from the first-person perspective.

So far I have suggested an empirical approach to understanding the sense of self-as-object. The sense of full-body ownership provides theoretical and experiential grounds for understanding the sense of self-as-physical-body, which, in turn, helps to explicate the sense of self-as-object. This means that we can understand consciousness of self-as-object by studying the sense of full-body ownership. This fits Wittgenstein’s and Shoemaker’s assertions that the “I”-as object allows misrepresentation. The main advantage of my approach, however, lies in the fact that we know how to conduct empirical research on the sense of self-as-object. Now, in cognitive neuroscience there are plenty of exciting studies on full-body illusions and their neural mechanisms (Lenggenhager et al. 2007; Petkova & Ehrsson 2008; Ehrsson 2007; Ehrsson 2012; Ionta et al. 2011; Blanke 2012; Serino et al. 2013). A philosophical account will certainly benefit from looking at these. But what about the sense of self-as-subject? In the next section, I will appeal to the notion of experiential ownership in order to capture this basic form of self-consciousness.

3 Experiential ownership and self-as-subject

The sense of experiential ownership is not about ownership of body parts or a whole body, but about whether I represent myself as the unique *subject* of experience. As I am typing, for example, I do not only experience tactile sensations in my fingers. I also have a sense that I am the one who is having these tactile sensations. This corresponds to Wittgenstein’s assertion: “To ask ‘are you sure it is you who have pains?’ would be nonsensical.” In this section, I will (1) illustrate that the sense of experiential ownership is different from the sense of body ownership; and (2) draw two distinctions to explicate the sense of experiential ownership. I will then (3) describe some varieties of the immunity principle (IEM); and (4) provide two counterexamples against two major forms of

IEM. We will see that, *pace* Wittgenstein and Shoemaker, we need another way of articulating the distinction between the sense of self-as-object and the sense of self-as-subject.

Moro et al. (2004) describe two patients with somatoparaphrenia. These patients suffered not only from somatoparaphrenia but also from hemispatial neglect and tactile extinction. They denied ownership of their left hand, in which they had no sensation, and their left visual field was lost. So far, we might think that these cases involve only misrepresentation of body ownership. But there is more. When the researcher moved the patients' left hand to the right-hand side so that they could see it, their tactile sensation was restored. But despite representing themselves as the subjects who felt the sensations, the two patients still denied the ownership of their left hands (2004, p. 440–441). This shows that it is possible to have the sense of experiential ownership without the sense of body ownership. The two types of self-experience are conceptually and empirically dissociable.

To clarify the notion of experiential ownership, let me begin with the point that every phenomenal state has a *what*-component and a *who*-component. The *what*-component includes the representational content and the phenomenal character of that state. The *who*-component ties the *what*-component to a unique subject. The basic assumption here is that every phenomenal state has one and only one subject. The sense of experiential ownership is exclusively about the *who*-component—it concerns whether one experiences oneself as the subject of a phenomenal state. I will now draw two distinctions to further clarify this point.

The first distinction is between the *fact* of experiential ownership and the *sense* of experiential ownership. When a subject experiences a phenomenal state, there exists a *fact* that he is the subject of that state. This fact of experiential ownership is constitutive of every conscious experience—i.e. every experience has a unique subject. For every conscious experience, we can ask “Who is the subject of that experience?” and there exists a fact of the matter. For example, right now it is me, not you, who is ex-

periencing lower-back pain. The fact of experiential ownership is objective in that it refers to a biological fact about whether a subject undergoes a phenomenal state.

When a subject experiences herself *as* the unique subject of a phenomenal state, she has the *sense* of experiential ownership, i.e. she experiences herself as the subject of that state. This aspect is captured by the Wittgenstein Question: “Are you sure it is you *who* has pains?” When a subject answers this question, she relies on her sense of experiential ownership. When I have a tactile sensation, I experience *what it is like for me* to undergo that sensation. The *what-it-is-like* aspect, i.e., the phenomenal character, belongs to the *what*-component. The *for-me* aspect refers to the subjective sense that I am the one who is having the sensation.¹³

The fact of experiential ownership and the sense of experiential ownership are two different aspects of experiential ownership: the *factual* aspect and the *subjective* aspect. These are not numerically different states or events that can be detached from a phenomenal state. Rather, they are two ways of characterizing the *who*-component of that state. The factual aspect addresses whether a subject experiences a phenomenal state; the subjective aspect concerns whether the subject is conscious of the factual aspect. But many philosophers do not see that these two aspects are not the same. To sustain this distinction, I will later argue that the factual aspect of experiential ownership can be misrepresented, which means that sometimes the Wittgenstein Question can be perfectly intelligible. Misrepresentation, as I shall explain, happens when the subjective aspect fails to match the factual aspect of experiential ownership.

The second distinction is between the first-personal sense and third-personal sense of experiential ownership. Suppose I experience a phenomenal state—say, lower-back pain. Not only do I experience the phenomenal character of the pain but also, *in the very same experience*, I have the sense that it is *me* who is experiencing that particular pain. This sense of

¹³ For other views about the *for-me* aspect, cf. Kriegel (2009) and LeGrand (2007).

experiential ownership is first-personal, since it is part of the pain experience rather than resulting from a separate act of reflection. I experience a sense of experiential ownership by experiencing the pain without requiring any further attention or introspection.

Now suppose I participate in an experiment where several subjects receive tactile stimulations in a random order and everyone is simultaneously scanned with fMRI equipment.¹⁴ Later, using the fMRI data on my somatosensory cortex, I can judge whether it was me who experienced a particular stimulation a few minutes ago. In this case, the sense of experiential ownership is considered from the third-person point of view, where the sense of experiential ownership is the content of a further judgment or reflection rather than an integral part of the respective phenomenal states.

I suggest that the sense of self-as-subject is captured by the first-personal sense of experiential ownership. Being conscious of oneself-as-subject just *is* to experience oneself as the subject of a phenomenal state. This implies that the sense of self-as-subject is exclusively about the *who*-component of a phenomenal state—no parts of the *what*-component belong to it. The sense of self-as-subject concerns whether I experience myself as the subject of a phenomenal state and nothing else. For the rest of this paper, I will use the term “the sense of experiential ownership” strictly in the first-personal sense.

Can one’s sense of self-as-subject go wrong? Following Wittgenstein and Shoemaker, most philosophers believe that the answer to this question is negative. According to Shoemaker, “in being aware that one feels pain one is, tautologically, aware, not simply that the attribute *feel(s) pain* is instantiated, but that it is instantiated *in oneself*” (1968, pp. 563–564; emphases in original). Hence, when I consciously feel a sensation, I *cannot be wrong* about whether it is me who feels it. This immunity (IEM) is widely considered to be a conceptual truth.¹⁵

I want to argue, however, that both Shoemaker and Wittgenstein are wrong. IEM is not a conceptual truth, and sometimes it makes perfect sense to ask the Wittgenstein Question—namely, “Are you sure it is you who is having a so-and-so experience?” Using my own terms, I will argue that the *sense* of experiential ownership can misrepresent the *fact* of experiential ownership. First, let me briefly mention some varieties of IEM. (1) Pryor (1999) distinguishes between *de re* misidentification and *which-object* misidentification.¹⁶ *De re* misidentification is false identification of two particular objects. It occurs when a mental state that *a* is F involves an assumption that *a* = *b*, but in fact *a* ≠ *b*. For example, when looking in the mirror, I misidentify someone else as myself (Pryor 1999, p. 276). A mental state enjoys *de re* immunity just in case it is not possible for the state to be in error through *de re* misidentification. In the case of *which-object* misidentification, one makes an existential generalization that there is something that is F based on suitable grounds, but misidentifies *which thing* is F (Pryor 1999, p. 281). For example, when listening to a symphony orchestra, I can tell that one of the trumpet players is slightly out of tune, but I misidentify which one it is. A mental state enjoys *which-object* immunity just in case it is not possible for the state to be in error through *which-object* misidentification. (2) De Vignemont (2012) recently distinguished bodily immunity from mental immunity. Mental immunity concerns whether certain self-ascriptions of mental states, including thoughts, judgments, or sensations, etc., enjoy IEM. By contrast, bodily immunity is not about mental states but about bodily properties. It concerns whether certain self-ascriptions of bodily states enjoy IEM, e.g. “my legs are crossed”.¹⁷

the self requires that when ascribing a mental state to oneself, e.g. “*a* is F”, one needs to demonstrate both “*b* is F” and “*a* = *b*.” But “*b* is F” would in turn require both “*c* is F” and “*b* = *c*”, and hence generates an infinite regress. This, Shoemaker argues, shows that the sense of self-as-subject must be identification-free.

¹⁶ Although disputed (Coliva 2006), many still consider this distinction useful.

¹⁷ Other varieties of IEM have been proposed in the literature. For example, Shoemaker (1968) distinguishes between circumstantial and absolute immunity, and between *de facto* and logical immunity (Shoemaker 1970; cf. also Coliva 2006). Pryor (1999) distinguishes between relative and absolute immunity. The former refers to im-

¹⁴ The method used here is called hyperscanning; cf. Montague et al. (2002).

¹⁵ Also, when specifying the “I”-as-subject, Shoemaker remarks that “not every self-ascription could be grounded on an identification of a presented object as oneself” (1968, p. 561). Because identification of

My target is a form of mental immunity that I call *experiential* immunity. *Experiential* immunity concerns phenomenal experiences. It is a form of *relative* immunity—that is, it is relative to first-personal access to phenomenal states, such as introspection, somatosensation, proprioception, etc. Experiential immunity is then the phenomenon that, when I am aware of a phenomenal state through first-personal access, I cannot be wrong about whether it is me who feels it. Experiential immunity can be construed as *de re* or *which-object* immunity. In the following section, I present counterexamples against both versions of experiential immunity. This will show that the sense of self-as-subject can be erroneous.

Bottini et al. (2002) describe a somatopraphrenia patient (“FB”) who has lost tactile sensation in her left hand and insists that her left hand belongs to her niece. They conducted the following tests on the patient, each involving several trials: (i) FB was blindfolded and told by the researcher that her left hand would be touched. Then the researcher actually touched the dorsal surface of her left hand. The result was that FB always reported feeling no sensation. (ii) FB was again blindfolded and was told that her *niece’s* hand would be touched. The result in this case was that, when the researcher touched the dorsal surface of her left hand, surprisingly, FB reported feeling the touch.¹⁸ The relevance of this case to IEM lies in the fact that, since FB was blindfolded during these tests, she relied on internal and first-personal access (e.g., introspection, somatosensation, proprioception) to determine whether or not she felt the touch. The perplexity lies in the difference between tests (i) and (ii). For the researcher, the only difference between the two

was the verbal cues given to FB before touching her hand. The remaining conditions were the same. But for FB, the difference was dramatic. Why is it that FB felt nothing when she expected that she herself would be touched, but felt the sensations when she expected that her niece would be touched? What is the best description of this strange phenomenology?

My view is that, during test (ii), FB misrepresented her tactile sensations as belonging to someone else, namely her niece. For the sake of argument, Shoemaker and I can agree on the following claims: (1) for every phenomenal state there must be a subject who experiences it; (2) every phenomenal state is in principle available to first-personal access (Shoemaker 1996); (3) every phenomenal state is experienced by the one who has first-personal access to that state. The crucial point is that (1)–(3) do not imply that (4) every phenomenal state is, from the first-person point of view, *represented as* experienced by the one who has first-personal access to that state. In FB’s case, (4) fails. FB fails to represent from her first-person perspective that she is the owner of the sensations. During test (ii), the factual aspect of her experiential ownership of the tactile sensations was intact when she was told that her niece would be touched, i.e., she was indeed the one who felt the tactile sensations. What went wrong was her sense of experiential ownership. Although FB felt the sensations, she misrepresented this fact as it being her niece who felt them.¹⁹ This shows that it is empirically possible for a subject, while being aware of a phenomenal state via a first-personal

munity relative to certain rational grounds G, and the latter immunity by every possible ground. Regarding judgments and beliefs, Coliva (2006) suggests a distinction between immunity relative to the subject’s own rational grounds and immunity relative to background presuppositions.

¹⁸ Test (ii) was conducted for four sessions, and FB reported feeling touches in 70%, 70%, 100%, and 80% of the trials respectively. As Bottini et al. observe: “her tactile imperceptions dramatically recovered” (2002, p. 251). To test if FB was just guessing, she was again blindfolded and told that her right hand (which is normal) would be touched. But actually the researcher did not touch her right hand. The result was that FB never reported feeling sensations—i.e., she passed the catch trials.

¹⁹ Shoemaker describes IEM as follows: ‘The statement ‘I feel pain’ is not subject to error through misidentification relative to ‘I’: it cannot happen that I am mistaken in saying ‘I feel pain’ because, although I do know of someone that feels pain, I am mistaken in thinking that person to be myself’ (1968, p. 557). Based on this description, some might insist that the self-ascriptions involved in IEM must be propositional in form, i.e. judgments, beliefs or statements. However, I contend that this restriction is unnecessary. What is crucial for IEM is that the self-ascriptions are based on first-personal grounds such as introspection, somatosensation, and proprioception, etc. As Bottini et al. have stated: “The patient was blindfolded and instructed to say ‘yes’ when she felt a touch and ‘no’ when she did not feel any touch” (2002, p. 251). So when FB said “yes” during test (ii), there is no reason why this wouldn’t count as a self-ascription. Applying Shoemaker’s description to FB’s case: I am mistaken in reporting ‘yes’ during test (ii) because, although I do know of someone that feels the sensations (via first-personal access), I am mistaken in my thinking about who that person is. Shoemaker’s IEM can be violated.

method, to commit a *de re* error regarding who the subject of that state is. Hence, *de re* immunity fails. Using my own terms, the sense of experiential ownership can misrepresent the fact of experiential ownership.²⁰

The second case against Shoemaker's IEM is the "body swap illusion" (Petkova & Ehrsson 2008, figure 6). This involves agentive experience—I experience myself as someone who is doing something. In an experiment, subjects wore a head-mounted display (HMD), and stood face-to-face with the experimenter, who wore two closed-circuit television (CCTV) cameras. The images registered by the CCTV cameras were transmitted concurrently to the subjects' HMD, such that through the HMD the subjects saw their own body facing themselves. Both the subjects and the experimenter extended their right hands, took hold, and then squeezed synchronously for two minutes. Twenty college students participated in this experiment. The authors describe their phenomenology: "after the experiment, several of the participants spontaneously remarked: 'I was shaking hands with myself!'" (2008, p. 5)

This strange phenomenology indicates that the subjects' agentive experience was mistaken. It was the experimenter who was shaking their hands, not the subjects themselves. Again, Shoemaker and I can agree that: (1) for every agentive experience there must be a subject who experiences it; (2) every agentive experience is in principle available to first-personal access; and (3) every agentive experience is experienced by the subject who has first-personal access to it. However, (1)–(3) together do not imply that: (4) every agentive experience is, from the first-person perspective, *represented as* experienced by the subject who in fact has first-personal access to it. In this case, *which-object* immunity fails because (4) was violated by

those who experienced the strange phenomenology in the body swap illusion. They were aware that there was someone having the agentive experience of squeezing their hands, but they misrepresented themselves as the subject of that experience.²¹

As such, it is possible for the subject of a given conscious experience, while being aware of that experience via a first-personal standpoint, to be mistaken about who the subject is.²² Thus Wittgenstein is wrong: it would make perfect sense to ask FB and the body-swap subjects: 'Are you sure that it is *you* who is having a so-and-so experience?' And Shoemaker is wrong, too: experiential immunity is violated both in FB's case and in the body-swap illusion. One's sense self-as-subject can be mistaken—that is, the sense of experiential ownership can misrepresent the fact of experiential ownership. Therefore, since both the sense of self-as-object and the sense of self-as-subject can involve misrepresentation, Shoemaker's IEM fails to distinguish between them.²³

21 Again, one might wonder whether the misrepresentation in this case was about the judgment rather than about the sense of experiential ownership. My reply is that since the subjects were normal college students, their reportability was not in question. So it is plausible to assume that their reports that "I was shaking hands with myself" were based on their subjective phenomenology, and more specifically on their sense of experiential ownership. Hence it was their sense of experiential ownership that committed misrepresentation.

22 There are at least two other (possible) cases of misrepresentation of the sense of experiential ownership. One is *voice ownership*: an illusion in which a stranger's voice, when presented as the auditory concomitant of a participant's own speech, is perceived as a modified version of one's own voice. "It felt as if the voice I heard was my voice" (Zeng et al. 2011). The other is *perception ownership*: A twenty-three-year-old male (DP) suffered from right inferior temporal hypometabolism (Zahn et al. 2008). The authors of a study on this male described his sensations as follows: "It appeared to him that he was able to see everything normally, but that he did not immediately recognize that he was the one who perceives and that he needed a second step to become aware that he himself was the one who perceives the object."

23 Let me briefly compare my position with other views. First, following Shoemaker, Coliva (2000) states that "If a subject is introspectively aware of pain, this just means that she is feeling pain [...] it is a matter of *conceptual truth* that if a subject is introspectively aware of a certain mental state, then she herself is having it and, therefore, that *mental state is her own*" (my emphasis). In contrast to Coliva, my account rejects IEM as a conceptual truth. From the fact that a subject experiences a mental state it does not necessarily follow that the subject represents herself as the one who experiences that state. I take the possibility of misrepresentation to be an important feature of the sense of experiential ownership. Second, Legrand (2007) emphasizes that consciousness of self-as-subject is pre-reflective, meaning that it is not an object of intentional consciousness. She says that the self-as-subject "is neither an external object (for example, it is not my body that I can observe in

20 One might object that the mistake that FB made was about the judgment of experiential ownership, not the sense of experiential ownership. My reply is that since FB was blindfolded, her report was based on first-personal grounds, i.e. on introspection. In addition, FB passed the catch trials mentioned in. As Bottini et al. have stated, FB "did not show any other sign of mental deterioration on the Mini Mental State Examination" (2002, p. 251). Therefore, no evidence suggests that her reports were unreliable. These considerations support the idea that the mistake was FB's sense of experiential ownership rather than her judgment. For other objections and responses, cf. Lane & Liang (2011).

I propose that this distinction can be made clearer by looking again at the sense of body ownership and the sense of experiential ownership. As I suggested in the last section, the sense of self-as-object can be understood in terms of a sense of self-as-physical-body which, in turn, can be understood via a sense of full-body ownership. Hence, when one experiences full-body ownership, one is conscious of oneself-as-object. In this section, I have suggested that we take an empirical approach to understanding the sense of self-as-subject. We can understand the consciousness of self-as-subject by studying the sense of experiential ownership.²⁴ In the

the mirror) nor an internal object [...] I am simply looking outside at the external world, and within this single act of consciousness I pre-reflectively experience myself-as-subject" (2007). I agree that the sense of self-as-subject is often implicit rather than explicit. But Le-grand's view neglects the distinction that I draw between the fact and the sense of experiential ownership. This is indicated by the fact of her embracing IEM. The fact of experiential ownership can be secured simply by looking outside at the external world, but whether one's sense of experiential ownership is correct is another issue.

²⁴ What is the relationship between the sense of body ownership and the sense of experiential ownership? The short answer is that the former presupposes the latter, but a full treatment would require another paper. Here, let me draw on Metzinger's Self-model Theory of Subjectivity (2003, 2008) to briefly address this issue. According to this theory, PMIR (phenomenal model of the intentionality relation) is a phenomenal experience that represents the relation between a subject and an object component. For example, I take a bite of an apple. The PMIR contains a subject component (I), a relation component (tasting), and an object component (the apple). But I want to propose a revised version of PMIR. Since the PMIR is a complex phenomenal property experienced by a subject, it would sometimes be legitimate to ask who is experiencing this particular PMIR. Does the subject attribute the sense of experiential ownership of this PMIR to him or herself? My proposal is that PMIR consists of three components: (1) the sense of experiential ownership; (2) intentional relations; and (3) an object component. On this view, PMIR already involves the sense of experiential ownership as the subject component, which is distinct from intentional relations and the object component. This revised version of PMIR helps to unpack the phenomenological structures of the sense of body ownership as follows. The subject component is served by the sense of experiential ownership. The object component can be one of the following: my hand, a rubber hand, someone else's leg, my whole-body, or a virtual body, etc. The intentional relations include vision, touch, proprioception, location, motion, introspective awareness, affective feelings, and so on. Four quick remarks are relevant here. First, the sense of body ownership is itself a phenomenal state, about which (2) and (3) specify the *what*-component. The *who*-component of the sense of body ownership is characterized by (1) the sense of experiential ownership. Hence, the sense of body ownership presupposes the sense of experiential ownership. Second, it is (1) and (2) that generate the sense that (3) is part of my body. Third, the difference between the sense of body-part ownership and the sense of full-body ownership lies in (3), while (1) and (2) may remain the same. Finally, based on my proposal in section 1, the sense of self-as-physical-body can be understood in terms of the following structure of PMIR: (1) the sense of experiential ownership; (2) intentional relations; and (3) a whole body. And the sense of self-as-object can be understood in terms of the same structure of PMIR as well.

next two sections, I examine some of the most relevant empirical accounts about the sense of self-as-subject. I argue that none of them are satisfactory. The reasons for this will be valuable when we consider where to go from here.

4 Core-self and affective-self

For animals, many biological values, such as finding food and shelter, avoiding predators, etc., have to do *homeostasis*—namely maintaining overall physiological states within the range required for survival (Damasio 1999, 2010; Panksepp 1998, 2005). To explain this, both Damasio and Panksepp propose that the brain has distinctive emotion systems and self-systems (the “proto-self” and the “core-self”). These inter-connected systems regulate homeostasis by integrating external information from perception with internal information from the body.²⁵ Despite their differences, Damasio and Panksepp share the following views: (1) emotions and homeostasis play essential roles in explaining how the sense of self is generated in the brain; (2) the key brain areas related to the self involve not only cortical but also sub-cortical regions, especially the brain stem possessed by both humans and many animals; (3) those brain areas are crucial, because multifarious types of neural information are integrated in those regions and provide representations of the whole body; (4) both Damasio and Panksepp believe that their accounts explain not only the sense of self-as-object but also the sense of self-as-subject. In the following I elucidate these points and then examine whether their goals are achieved.

According to Damasio, animal brains have what he calls the proto-self system, which is “a dynamic collection of integrated neural processes, centered on the representation of the living body” (2010, p. 9). The neural processes of

²⁵ Both Damasio and Panksepp distinguish between emotions and their neural substrates, on the one hand, and feelings (Damasio) or affective feelings (Panksepp), on the other. Emotions refer to innate patterns of neural and physiological responses to environmental events. Feelings (or affective feelings) refer to phenomenal consciousness of emotions (Damasio 1999, p. 42, p. 55; Damasio 2010, pp. 108–110; Panksepp 1998, pp. 48–49; Panksepp 2005, p. 32). The emotion-systems closely interact with the self-systems to regulate and manage homeostasis.

this system represent “moment by moment, the most stable aspects of the organism’s physical structure”, on the one hand, and “the externally directed sensory portals”, on the other (2010, p. 190). This generates *primordial feelings* that “reflect the current state of the body” and “provide a direct experience of one’s own living body, wordless, unadorned, and connected to nothing but sheer existence” (2010, p. 21, p. 185). The proto-self system and primordial feelings account only for the sense of self-as-object (2010, p. 9, p. 202). The sense of self-as-subject is generated when an animal interacts with the environment such that a neural representation of the interaction is generated in the brain (2010, pp. 9–10, p. 91, p. 202). By interacting with external objects, the current state of the body and the proto-self system are modified. This modification activates the core-self system, which enhances attention to external objects and “engenders a sense of ownership” (2010, pp. 202–203). This is closely related to the sense of experiential ownership discussed above. It is part of what Damasio calls *core consciousness*, which “displays [...] moment by moment, that you rather than anyone else are doing the reading and the understanding of the text” (1999, p. 10).

Damasio’s key idea is that the brain produces not only first-order representations of external objects and of the body (2010, p. 76, p. 84, pp. 91–97), but also *second-order representations* of the relationship between objects and the organism (1999, pp. 169–170; 2010, pp. 71–72, p. 181). These are “the source of the sense of the self in the act of knowing” (1999, p. 169). When the core-self is *felt* (1999, p. 172), i.e. when the second-order representations become conscious states (2010, p. 248), core consciousness emerges. This includes a minimal sense of self-as-subject, a transient sense that “it is you [...] doing the seeing” (1999, p. 169; cf. 2010, p. 168), or the sense that I am the subject of current experiences (cf. 2010, p. 185, p. 203, p. 209). As we can see, this account is highly relevant to our current investigation.

Damasio emphasizes that the most crucial neural structures related to the proto-self and the core-self systems are found in the subcor-

tical regions, especially the brain stem (2010, p. 195, p. 205).²⁶ They include, among others, the nucleus tractus solitarius (NTS), the parabrachial nucleus (PBN), the periaqueductal gray (PAG), the hypothalamus, and the superior colliculus (2010, pp. 98–99, pp. 191–192; 1999, pp. 180–183). Why are these neural structures so critical for the core-self and core consciousness? According to Damasio, core consciousness results from *integration* of interoceptive, proprioceptive, and exteroceptive information, which produces second-order representations (2010, p. 76, p. 97, pp. 190–196, p. 199, p. 203, pp. 206–209). The brain areas just mentioned receive input from many other regions, which process information about external objects and internal bodily conditions (2010, p. 78, p. 80, pp. 84–85, p. 94, pp. 99–100, pp. 207–209). Thus it is in these areas that integration is thought to take place. Integration in those areas constitutes core consciousness because they provide neural representations of the organism’s *whole body* (2010, p. 68, pp. 94–97, p. 209, pp. 244–245), and the integration is implemented by neural synchrony in the gamma range (2010, p. 20, pp. 86–87).

Panksepp points out seven basic innate emotion-systems in mammals: seeking, rage, fear, lust, care, panic, and play.²⁷ These emotion-systems generate affective feelings, which characterize how animals respond to environmental challenges. Panksepp & Northoff (2009) also postulate that the proto- and core-self systems monitor and regulate homeostasis. The proto-self is ‘the most ancient form of coherent body representation’, and the core-self gives rise to “affective consciousness”.²⁸ Both systems are

²⁶ For Damasio, the cortical areas that are important for the core self include insular and somatosensory cortices (2010, pp. 205–209).

²⁷ According to Panksepp, emotions and affective feelings are internally generated by neuronal mechanisms to respond to life-challenging events. The neural systems of emotions compute and monitor homeostasis by evaluating an organism’s adaptation to the environment. Each emotion system refers to a specific neural network, mainly in the subcortical areas.

²⁸ Panksepp and Northoff prefer to use the expressions “proto-SELF” and “core-SELF” to emphasize neural mechanisms rather than mental phenomena, but this emphasis need not concern us here. They describe the relation between core-SELF and affective consciousness as follows: “What is subjectively experienced here is the relation of one’s body to the incentives in the environment as well as internally generated emotional arousals—the core-SELF thus enables the organism to access this relation in terms of subjective experience, e.g., a primitive form of phenomenal consciousness, which at this level is essentially affective” (2009, p. 196).

causally mediated by what they call affective *self-related processing*, which integrates interoceptive information from the body and exteroceptive stimuli from the environment. The main mechanism that underlies this processing is a subcortical-cortical midline system (SCMS) (2009, p. 197). The subcortical parts of this network include “the Periaqueductal gray (PAG), the superior colliculi (SC), and the adjacent mesencephalic locomotor region (MLR), as well as preoptic areas, the hypothalamus, and dorso-medial thalamus (DMT)” (2009, p. 201). On the superior colliculi (SC) and the periaqueductal gray (PAG), they tell us that:

The colliculi and the PAG are among the most richly connected areas of the brain; both receive afferents from several exteroceptive sensory regions (occipital, auditory, somatosensory, gustatory, and olfactory cortex) and, at the same time, afferents from other interoceptive subcortical regions. In addition, the PAG and the colliculi are connected with the cortical midline structures (CMS). (2009, p. 201)

Like Damasio, Panksepp and Northoff believe that the SC and the PAG play important roles in instigating the core-self system because they are the central areas where exteroceptive sensory information and interoceptive bodily information are integrated. They suggest that, due to anatomical convergence and neural synchronizations within the SCMS, “an archaic scheme of the *entire body* may be constituted in brain regions as low as the medial brainstem” (2009, p. 202; my emphasis).

Panksepp and Northoff claim that their theory explains what philosophers call the ‘experiential self’ and the ‘primitive form of selfhood’ (2009, p. 209). Self-related processing “intrinsically integrates affectivity, appropriateness and belongingness, and the phenomenal dimension of mineness into the *ownership of experience*” (2009, p. 199; my emphasis). This comes very close to the sense of experiential ownership that I discussed above. They consider self-related processing by the SCMS to be the mech-

anism not only of affective consciousness but also of the sense of self-as-subject.

In sum, Damasio, Panksepp and Northoff suggest that the sense of self-as-subject can be explained by full-body representations implemented by neural synchrony or by the SCMS. Now the key issue is: Do their accounts really specify the neural mechanisms that produce the sense of self-as-subject? Or do they specify only the mechanisms of the sense of self-as-object, i.e., of consciousness of oneself as a physical body interacting with the world? I argue that they address only the sense of self-as-object; they do not really provide a genuine account of the sense of self-as-subject.²⁹ Below I raise this theoretical issue; empirical arguments will follow in the next section.

Damasio claims that core consciousness is constituted by a second-order neural representation of the relation between animal and the environment. But this seems to require more explanation. Yet an explanation is not really provided by Damasio. I can agree that, for the sense of self-as-object, one must not only represent the external world, but also the body. But we cannot assume that the same account will automatically apply to the sense of self-as-subject. The problem with Damasio’s account is that the theoretical link between full-body representation and the sense of self-as-subject is lacking. And Panksepp and Northoff’s account is afflicted with the same defect. It might be that full-body representations are part of the biological conditions *necessary* for generating the sense of self-as-subject. But since they are also necessary for the sense of body ownership and the sense of self-as-object, it is far from obvious whether they are *sufficient* for the sense of self-as-subject. Let me elaborate.

Consider the full-body illusion mentioned in section 1. According to Blanke & Metzinger (2009), this illusion contains three central features related to self-consciousness. The first is *self-identification*. When the subjects experienced OBE during the experiment, “they felt as if the virtual body was their own” (2009, p. 12). We can see that this feature turns on the ques-

²⁹ Cf. Legrand (2007) for a slightly different criticism of Damasio.

tion “Is that body mine?” rather than “Am I the one who is having this experience?” So self-identification is about the sense of full-body ownership rather than the sense of experiential ownership. The second feature is *self-location*, which concerns “where my body is located in space and time”. Again, this is about the spatiotemporal position of the body rather than the sense of experiential ownership. Blanke and Metzinger call the third feature a *weak first-person perspective*, defined as a geometrical point of projection and nothing more (2009, p. 8). So construed, even a camera could possess such a perspective. Hence, this feature does not specify the sense of self-as-subject, either.

The point is that, in the OBE experiment, the sense of experiential ownership is not in question and hence not measured. This means that explanations of the mechanisms of full-body representation or the sense of body ownership do not necessarily apply to the sense of experiential ownership. As such, self-related processing can help explain full-body representation *without* explaining the sense of self-as-subject. Damasio, Panksepp and Northoff neglect the theoretical gap between full-body representation and the sense of self-as-subject, hence their accounts do not really explain the sense of self-as-subject. They suggest that the sense of self-as-subject results from integration by neural synchrony in the brain stem or the SCMS. But it remains unexplained why and how this could be so. To investigate these worries, I examine in the next section the two major proposals by neuroscientists regarding the mechanisms of the sense of self-as-subject: neural synchrony and processing in the SCMS.

5 Neural synchrony and subcortical-cortical midline structures

Neurons in different brain regions may exhibit rhythmic firing patterns. This is called neural oscillation, the frequency of which can be recorded by an electroencephalogram (EEG). When a group of neurons fire together with the same oscillation pattern, they are in *synchrony*. Neural synchrony is considered to be a central mechanism of many cognitive functions. In the

case of conscious perception, multifarious types of visual information are processed in different brain regions, which need to be combined in order to produce coherent percepts. Many researchers suggest that transient synchronization in the visual system provides such a binding mechanism (Engel & Singer 2001; Singer 2004; Singer 2007; Koch 2004). In addition to vision, synchronization in the beta and gamma ranges is also found in the olfactory, auditory, and somatosensory systems, as well as in other brain areas that influence (or are influenced by) perception, such as the pre-frontal cortex, the motor cortex, and the hippocampus (Singer 2007).

However, if this is all there is to neural synchrony, it would not explain the sense of self-as-subject at all. What we are looking for is not the mechanism that explains what I consciously perceive, but the mechanism that produces the sense that I, rather than someone else, am the subject of these perceptions.³⁰ Thus, information integration by neural synchrony may explain the content of consciousness without explaining the sense of experiential ownership, i.e., it explains *what* one experiences rather than *who* the subject of that experience is. In the following I consider three recent developments that connect neural synchrony more closely with self-consciousness.

(1) Uhlhaas et al. (2009) recently suggested that there are high correlations between disorders of self-consciousness and abnormalities in neural synchrony. Symptoms of schizophrenia, epilepsy, autism, Alzheimer’s disease, and Parkinson’s disease are related to dysfunctions of synchronization. For example, correlations have been suggested between reduced or abnormal alpha- or gamma-band oscillations, on the one hand, and impaired visual binding, auditory hallucination in schizophrenia, and impaired linguistic and auditory performance in autism, on the other. The problem is that the sense of experiential ownership is not itself targeted in these studies. Researchers measured how abnormal neural synchrony relates to impaired cognitive performance, rather than to who the subject of the experience is.

³⁰ The sense of experiential ownership is not studied in Singer’s work on neural synchrony at all.

(2) [Lou et al. \(2010\)](#) used transcranial magnetic stimulation (TMS) to show that a medial paralimbic network is crucial for minimal self-consciousness.³¹ This network may “bind conscious experiences with different degrees of self-reference through synchrony of high frequency oscillations” (2010, p. 185). They tested three conditions that represent different degrees of self-reference: maximal (“Self”), intermediate (“Franz”), and minimal (“Syl”). In each condition a set of adjectives were sequentially presented on a screen.³² In the “Self” condition, the subject’s task was to make personal judgments concerning how well each adjective fitted him or herself. However, none of these conditions are about the sense of experiential ownership. Whether it was “I” who looked at the screen and made the judgments was not in question. Hence, the sense of self-as-subject was not measured by the reported patterns of synchronization.

(3) [Kanayama et al. \(2009\)](#) used EEG to investigate the rubber hand illusion (RHI), and found high correlation between the visual-tactile integration process and gamma-band synchrony in the parietal cortex. The stronger the subjects experienced the illusion, the higher the synchrony was. The authors suggested that RHI is caused by gamma band synchrony. In addition, a study of the full-body illusion by [Lenggenhager et al. \(2011\)](#) found high correlation between alpha-band oscillations in the sensorimotor cortex and the medial prefrontal cortex, on the one hand, and subjects feeling themselves to be located in space, on the other. Unfortunately, these studies do not really tell us about the sense of self-as-subject. In these experiments, what was misrepresented was the sense of ownership of a body part or a whole body. Whether “I” was the one who was experiencing the illusions was not in question. The synchronization reported by these studies can help explain the sense of body ownership, but not the sense of self-as-subject.

³¹ [Lou et al. \(2010\)](#) suggest that this network includes the anterior cingulate, medial prefrontal and posterior cingulate, and the medial parietal cortices, connected via the thalamus.

³² In the ‘Franz’ condition, the subject judged how well each adjective fitted a well-known German football star Franz Beckenbauer. In the “Syl” condition, the subject’s task was to decide whether each of the different sets of adjectives had an even or odd number of syllables.

As far as I know, no empirical study on neural synchrony really targets the sense of self-as-subject. We cannot explain the sense of experiential ownership simply by describing the mechanisms of content of conscious perception, cognitive deficits, or body ownership. The lesson here is that we need first to ascertain that the neural information being integrated by synchrony is *about* the sense of self-as-subject, and not just about representation of the organism’s bodily condition. Unless we know exactly how the integrating processes bring about that one represents oneself as the subject of phenomenal or conscious states, we cannot say that the mechanisms of the sense of self-as-subject have been found. As I will suggest below, the key here is to identify the right research question. And this is where philosophy can make contributions to neuroscience.

The second proposal regarding the mechanisms of the sense of self-as-subject, suggested by [Panksepp & Northoff \(2009\)](#), is self-related processing implemented in the subcortical-cortical midline system (SCMS). This mechanism is notably related to the so-called resting state and the default mode network. Researchers have found that some brain areas are highly activated in the resting state, i.e. when the subject is not actively engaging with its environment (e.g. lying quietly in a scanner with eyes closed but awake) ([Raichle et al. 2001](#)). Interestingly, the activations decrease significantly when the subject performs tasks that involve focusing on the external world. These brain areas constitute what is now called the default mode network.

How one should interpret the neural activities in the resting state and the default mode network, and how they relate to self-consciousness, are controversial issues. For example, [Gillihan & Farah \(2005\)](#) point out that different research programs on the self employ divergent methodologies and implicate a wide range of brain areas. Putting all the data together, we do not obtain a specific or unitary picture, because pretty much the entire brain is involved in processing the sense of self. This and other criticisms suggest that we should be

cautious when interpreting the alleged empirical evidence about the sense of self-as-subject.³³

Still, many researchers maintain that resting state activities and the default mode network are closely related to the self (cf. Gusnard 2005; D'Argembeau et al. 2007). Northhoff et al. (2006) reviewed a vast number of imaging studies, and compared the processing of what they call self-related tasks and non-self-related tasks.³⁴ They found that the data indicate the same group of brain areas, including “the medial orbital prefrontal cortex (MOFC), the ventromedial prefrontal cortex (VMPFC), the sub/pre- and supragenual anterior cingulate cortex (PACC, SACC), the dorsomedial prefrontal cortex (DMPFC), the medial parietal cortex (MPC), the posterior cingulate cortex (PCC), and the retrosplenial cortex (RSC)” (2006, pp. 441–442). These areas constitute the cortical midline structures (CMS), i.e. the cortical parts of the SCMS. Compared with non-self-related tasks, when subjects perform self-related tasks their CMS reveal high activation across all domains (2006, p. 450). The authors suggest that the CMS correspond to the default mode network,³⁵ and that neural activity in the CMS constitutes “an experiential self that mediates ownership of experience” (2006, p. 441). “Ownership”, they claim, “describes the sense that I am the one who is undergoing an experience” (2006, p. 448), which makes this account directly relevant to our investigation.

Legrand & Ruby (2009) argue against Northhoff et al. that the CMS are at most self-related, i.e. related to the self only to some extent, but not *self-specific*, i.e., not specific

enough to capture the sense of self-as-subject.³⁶ Partly because of this criticism, but more because of new findings by his own group, Northhoff's view has changed significantly in recent times. First, Qin et al. (2010) recently studied the CMS in patients who are in a vegetative state. Surprisingly, by showing the patients their own names, various regions in their CMS were activated. Assuming that vegetative patients have lost the capacity to experience themselves as subjects, this finding undermines Northhoff's previous claim that the CMS constitutes an “experiential self that mediates ownership of experience.” In fact, Northhoff now agrees that the neural processing in the CMS is at most a necessary condition for the experiential self.³⁷

Second, after conducting a meta-analysis on eighty-seven imaging studies covering 1433 participants, Qin & Northhoff (2011) suggest that self-related processing involves far fewer areas in the CMS. It is the perigenual anterior cingulate cortex (PACC), rather than the medial prefrontal cortex (MPFC) or posterior cingulate cortex (PCC), that is specifically involved in self-processing. This indicates that they have become more cautious about interpreting data. However, they still maintain that there exists a strong connection between the PACC and the sense of self. They argue that “our sense of self may result from a specific kind of interaction between resting state activity and stimulus-induced activity, i.e., rest-stimulus interaction, within the midline regions” (2011, p. 1221). That is, a narrower network *within* the CMS is not just necessary but indeed sufficient for “generating our sense of the self” (2011, p. 1222). I will comment on this last claim below.

Whether or not Qin and Northhoff take their notion of “sense of self” to include the sense of

³³ Another criticism is that, when the subject is interacting with the world, the neural activity in the default mode network is not totally extinguished. Some studies show that it is “reorganized in response to the working memory task” (Fransson 2006). Others have suggested that it could “function to support exploratory monitoring of the external environment when focused attention is relaxed” (Buckner et al. 2008).

³⁴ Many of these studies used a “judgment paradigm”. Subjects made explicit evaluative judgments about first- vs. third-person perspectives, own vs. others' judgments, self vs. others' decisions, own vs. others' personality traits, etc. The domains that Northhoff et al. (2006) reviewed include verbal, spatial, memory, emotional, facial, agency, ownership of movements, and social tasks.

³⁵ CMS show a high level of neural activity during the resting state. Non-self-referential tasks elicit large signal decreases in the CMS (Northhoff et al. 2006, p. 450).

³⁶ Legrand and Ruby indicated that the CMS are involved not only in self-related tasks, but also in several cognitive tasks that are not related to self-consciousness at all. For example, their review showed that some areas in the CMS are activated in others' mind reading, inductive and deductive reasoning, resting state, and memory recall. Moreover, these areas are “sometimes more activated for the self than for others and sometimes more activated for others than for self” (Legrand & Ruby 2009, p. 258).

³⁷ Northhoff et al. tell us that “the neural mechanisms underlying SRP [self-related processing] may only be considered a necessary condition which is not sufficient by itself to constitute a self with its self-specific contents” (2011, p. 55).

self-as-subject, I argue that their meta-analysis does not capture the sense of self-as-subject. They describe the operational criteria as follows: “the specificity of the self (e.g. hearing one’s own name, seeing one’s own face) was tested and compared across familiar (using stimuli from personally known people) and other (non-self–non-familiar, i.e. strangers and widely-known figures) conditions” (2011, p. 1211). The tasks in the “self condition” include “trait adjective judgment, retrieval of personality traits, face recognition, body recognition, personal thinking, name perception, autobiographical memory, own feeling, self-administered pain, person perspective tasks and agency tasks” (2011, p. 1224). All these tasks are about participants making judgments about whether a certain property may be suitably attributed to themselves. From the first-person point of view, the participants are judging whether the contents of the stimuli accurately characterize themselves. But again, whether “I” am the one who is experiencing the stimuli and making the judgments is really not in question, and hence not reflected in the data. Once again, the sense of self-as-subject is not measured by Qin and Northoff’s most recent study.

I conclude that Damasio, Panksepp, and Northoff have all failed to explain the mechanisms of the sense of self-as-subject. A theoretical gap exists between neural synchrony and the SCMS, on the one hand, and the sense of self-as-subject, on the other. But it is important to see exactly where the shortcoming is. It is not that neural synchrony and the SCMS are completely irrelevant to the sense of self-as-subject. Rather, the failure is that why and how they are relevant have not really been explained. This is because the neuroscientists have not clarified and captured the sense of self-as-subject well enough, such that they over-interpret data and make unjustified claims about this target phenomenon. In this regard, my proposals in sections 2 and 3 have provided the required clarification.

6 Conclusion

I have suggested that the sense of self-as-subject can be explicated by examining the sense of experiential ownership, which is distinct from the

sense of body ownership. Having a conscious experience secures only the fact of experiential ownership, not the sense of experiential ownership. This provides a reinterpretation of the distinction between the sense of self-as-object and the sense of self-as-subject. I elucidated the sense of self-as-object by looking at the sense of body ownership, and the sense of self-as-subject by examining the sense of experiential ownership. It became clear that both can misrepresent. The possibility of misrepresentation makes the sense of self-as-subject open to empirical as well as philosophical investigations. It is important to investigate how misrepresentation of the sense of experiential ownership is generated. This requires us to identify the right research question—which, I suggest, is precisely the Wittgenstein Question. When examining pathological cases or conducting experiments, researchers should ask their subjects questions like: “Are you sure it is you who is feeling your niece’s sensations?” or “Are you sure it is you who is shaking your own hand?” Then psychophysical and fMRI experiments can be developed to study the subjects’ responses. As such, to move forward, the first step is to look for and then to study the various conditions about which one can pursue the Wittgenstein Question.

7 Acknowledgements

I wish to thank Thomas Metzinger, Jennifer Windt, Shaun Gallagher, Sascha Fink, Krisztina Orban, Jakob Hohwy, and Bigna Lenggenhager for their comments on this paper.

References

- Blanke, O. (2012). Multisensory brain mechanisms of bodily self-consciousness. *Nature Reviews Neuroscience*, 13, 556-571. [10.1038/nrn3292](https://doi.org/10.1038/nrn3292)
- Blanke, O. & Metzinger, T. (2009). Full-body illusions and minimal phenomenal selfhood. *Trends in Cognitive Sciences*, 13 (1), 7-13. [10.1016/j.tics.2008.10.003](https://doi.org/10.1016/j.tics.2008.10.003)
- Bottini, G., Bisiach, E., Sterzi, R. & Vallar, G. (2002). Feeling touches in someone else's hand. *NeuroReport*, 13 (2), 249-252. [10.1097/00001756-200202110-00015](https://doi.org/10.1097/00001756-200202110-00015)
- Botvinick, M. & Cohen, J. (1998). Rubber hands 'feel' touch that eyes see. *Nature*, 391 (6669), 756-756. [10.1038/35784](https://doi.org/10.1038/35784)
- Buckner, R., Andrews-Hanna, J. & Schacter, D. (2008). The brain's default network: Anatomy, function, and relevance to disease. *Annals of New York Academy of Sciences*, 1124, 1-38. [10.1196/annals.1440.011](https://doi.org/10.1196/annals.1440.011)
- Coliva, A. (2000). Thought insertion and immunity to error through misidentification. *Philosophy, Psychiatry, and Psychology*, 9 (1), 27-34. [10.1353/ppp.2003.0004](https://doi.org/10.1353/ppp.2003.0004)
- (2006). Error through misidentification: Some varieties. *Journal of Philosophy*, 103 (8), 403-425.
- Damasio, A. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. San Diego, CA: Harcourt.
- (2010). *Self comes to mind: Constructing the conscious brain*. New York, NY: Pantheon.
- D'Argembeau, A., Ruby, P., Collette, F., Degueldre, C., Baetee, E., Luxen, A., Maquet, P. & Salmon, E. (2007). Distinct regions of the medial prefrontal cortex are associated with self-referential processing and perspective taking. *Journal of Cognitive Neuroscience*, 19 (6), 935-944. [10.1162/jocn.2007.19.6.935](https://doi.org/10.1162/jocn.2007.19.6.935)
- de Vignemont, F. (2011). Bodily awareness. *Stanford encyclopedia of philosophy*. <http://plato.stanford.edu/archives/fall2011/entries/bodily-awareness/>
- (2012). Bodily immunity to error. In S. Prosser & R. Recanati (Eds.) *Immunity to error through misidentification: New essays* (pp. 224-246). Cambridge, UK: Cambridge University Press.
- Dokic, J. (2003). The sense of ownership: An analogy between sensation and action. In J. Roessler & N. Eilan (Eds.) *Agency and self-awareness: Issues in philosophy and psychology* (pp. 321-344). Oxford, UK: Oxford University Press.
- Ehrsson, H. (2007). The experimental induction of out-of-body experiences. *Science*, 317 (5841), 1048-1048. [10.1126/science.1142175](https://doi.org/10.1126/science.1142175)
- (2012). The concept of body ownership and its relation to multisensory integration. In B. Stein (Ed.) *The new handbook of multisensory processing* (pp. 775-792). Cambridge, MA: MIT Press.
- Engel, A. & Singer, W. (2001). Temporal binding and the neural correlates of sensory awareness. *Trends in Cognitive Sciences*, 5 (1), 16-25. [10.1016/S1364-6613\(00\)01568-0](https://doi.org/10.1016/S1364-6613(00)01568-0)
- Fransson, P. (2006). How default is the default mode of brain function? Further evidence from intrinsic BOLD signal fluctuations. *Neuropsychologia*, 44 (14), 2836-2845. [10.1016/j.neuropsychologia.2006.06.017](https://doi.org/10.1016/j.neuropsychologia.2006.06.017)
- Gallagher, S. (2000). Philosophical conceptions of the self: Implications for cognitive science. *Trends in Cognitive Sciences*, 4 (1), 4-21. [10.1016/S1364-6613\(99\)01417-5](https://doi.org/10.1016/S1364-6613(99)01417-5)
- (2005). *How body shapes the mind*. Oxford, UK: Oxford University Press.
- Gillihan, S. & Farah, M. (2005). Is self special? A critical review of evidence from experimental psychology and cognitive neuroscience. *Psychological Bulletin*, 131 (1), 76-97. [10.1037/0033-2909.131.1.76](https://doi.org/10.1037/0033-2909.131.1.76)
- Gusnard, D. (2005). Being a self: Considerations from functional imaging. *Consciousness and Cognition*, 14 (4), 679-697. [10.1016/j.concog.2005.04.004](https://doi.org/10.1016/j.concog.2005.04.004)
- Ionta, S., Gassert, R. & Blanke, O. (2011). Multi-sensory and sensorimotor foundation of bodily self-consciousness - An interdisciplinary approach. *Frontiers in Psychology*, 2, 1-8. [10.3389/fpsyg.2011.00383](https://doi.org/10.3389/fpsyg.2011.00383)
- Kanayama, N., Sato, A. & Ohira, H. (2009). The role of gamma band oscillations and synchrony on rubber hand illusion and crossmodal integration. *Brain and Cognition*, 69 (1), 19-29. [10.1016/j.bandc.2008.05.001](https://doi.org/10.1016/j.bandc.2008.05.001)
- Koch, C. (2004). *The quest for consciousness: A neurological approach*. Englewood, CO: Roberts and Company Publishers.
- Kriegel, U. (2009). *Subjective consciousness*. Oxford, UK: Oxford University Press.
- Lane, T. & Liang, C. (2011). Self-consciousness and immunity. *Journal of Philosophy*, 108 (2), 78-99.
- Legrand, D. (2006). The bodily self: The sensori-motor roots of pre-reflective self-consciousness. *Phenomenology and the Cognitive Sciences*, 5 (1), 89-118. [10.1007/s11097-005-9015-6](https://doi.org/10.1007/s11097-005-9015-6)
- (2007). Pre-reflective self-as-subject from experiential and empirical perspectives. *Consciousness and Cognition*, 16 (3), 583-599. [10.1016/j.concog.2007.04.002](https://doi.org/10.1016/j.concog.2007.04.002)
- (2010). Myself with no body? body, bodily-con-

- consciousness and self-consciousness. In S. Gallagher & D. Schmicking (Eds.) *Handbook of phenomenology and cognitive science* (pp. 181-200). Dordrecht, NL: Springer.
- (2011). Phenomenological dimensions of bodily self-consciousness. In S. Gallagher (Ed.) *Oxford handbook of the self* (pp. 204-227). Oxford, UK: Oxford University Press.
- Legrand, D. & Ruby, P. (2009). What is self-specific? Theoretical investigation and critical review of neuroimaging results. *Psychological Review*, 116 (1), 252-282. [10.1037/a0014172](https://doi.org/10.1037/a0014172)
- Lenggenhager, B., Tadi, T., Metzinger, T. & Blanke, O. (2007). Video ergo sum: Manipulating bodily self-consciousness. *Science*, 317 (5841), 1096-1099. [10.1126/science.1143439](https://doi.org/10.1126/science.1143439)
- Lenggenhager, B., Halje, P. & Blanke, O. (2011). Alpha and oscillations correlate with illusory self-location induced by virtual reality. *European Journal of Neuroscience*, 33 (10), 1935-1943. [10.1111/j.1460-9568.2011.07647.x](https://doi.org/10.1111/j.1460-9568.2011.07647.x)
- Lou, H., Gross, J., Biermann-Ruben, K., Kjaer, T. & Schnitzler, A. (2010). Coherence in consciousness: Paralimbic gamma synchrony of self-reference links conscious experiences. *Human Brain Mapping*, 31 (2), 185-192. [10.1002/hbm.20855](https://doi.org/10.1002/hbm.20855)
- Metzinger, T. (2003). *Being no one*. Cambridge, MA: MIT Press.
- (2008). Empirical perspectives from the self-model theory of subjectivity: A brief summary with examples. In R. Banerjee & B. Chakrabarti (Eds.) *Progress in brain research, Vol. 168, Models of brain and mind: Physical, computational and psychological approaches* (pp. 215-246). Amsterdam, NL: Elsevier.
- Montague, P., Berns, G., Cohen, J., McClure, S., Pagnoni, G., Dhamala, M., Wiest, M., Karpov, I., King, R., Apple, N. & Fisher, R. (2002). Hyperscanning: Simultaneous fMRI during linked social interactions. *NeuroImage*, 16 (4), 1159-1164. [10.1006/nimg.2002.1150](https://doi.org/10.1006/nimg.2002.1150)
- Moro, V., Zampini, M. & Aglioti, S. (2004). Changes in spatial position of hands modify tactile extinction but not disownership of contralesional hand in two right brain-damaged patients. *Neurocase*, 10 (6), 437-443. [10.1080/13554790490894020](https://doi.org/10.1080/13554790490894020)
- Northoff, G., Heinzel, A., Greck, M., Bermpohl, F., Dobrowolny, H. & Panksepp, J. (2006). Self-referential processing in our brain - A meta-analysis of imaging studies of self. *NeuroImage*, 31 (1), 440-457. [10.1016/j.neuroimage.2005.12.002](https://doi.org/10.1016/j.neuroimage.2005.12.002)
- Northoff, G., Qin, P. & Feinberg, T. (2011). Brain imaging of the self - Conceptual, anatomical and methodological issues. *Consciousness and Cognition*, 20 (1), 52-63. [10.1016/j.concog.2010.09.011](https://doi.org/10.1016/j.concog.2010.09.011)
- Panksepp, J. (1998). *Affective neuroscience*. Oxford, UK: Oxford University Press.
- (2005). Affective consciousness: Core emotion feelings in animals and humans. *Consciousness and Cognition*, 14 (1), 30-80. [10.1016/j.concog.2004.10.004](https://doi.org/10.1016/j.concog.2004.10.004)
- Panksepp, J. & Northoff, G. (2009). The trans-species core SELF: The emergence of active cultural and neuro-ecological agents through self-related processing within subcortical-cortical midline networks. *Consciousness and Cognition*, 18 (1), 18,193-215. [10.1016/j.concog.2008.03.002](https://doi.org/10.1016/j.concog.2008.03.002)
- Petkova, V. & Ehrsson, H. (2008). If I were you: Perceptual illusion of body swapping. *PLoS One*, 3 (12), e3832-e3832. [10.1371/journal.pone.0003832](https://doi.org/10.1371/journal.pone.0003832)
- Pryor, J. (1999). Immunity to error through misidentification. *Philosophical Topics*, 26 (1-2), 271-304. [10.5840/philtopics1999261/246](https://doi.org/10.5840/philtopics1999261/246)
- Qin, P., Di, H., Liu, Y., Yu, S., Gong, Q., Duncan, N., Weng, X., Laureys, S. & Northoff, G. (2010). Anterior cingulate activity and the self in disorders of consciousness. *Human Brain Mapping*, 31 (12), 1993-2002. [10.1002/hbm.20989](https://doi.org/10.1002/hbm.20989)
- Qin, P. & Northoff, G. (2011). How is our self related to midline regions and the default-mode network? *NeuroImage*, 57 (3), 1221-1233. [10.1016/j.neuroimage.2011.05.028](https://doi.org/10.1016/j.neuroimage.2011.05.028)
- Raichle, M., MacLeod, A., Snyder, A., Powers, W., Gusnard, D. & Shulman, G. (2001). A default mode of brain function. *PNAS*, 98 (2), 676-682. [10.1073/pnas.98.2.676](https://doi.org/10.1073/pnas.98.2.676)
- Rohde, M., Di Luca, M. & Ernst, M. (2011). The rubber hand illusion: Feeling of ownership and proprioceptive drift do not go hand in hand. *PLoS One*, 6 (6), e21659-e21659. [10.1371/journal.pone.0021659](https://doi.org/10.1371/journal.pone.0021659)
- Serino, A., Alsmith, A., Costantini, M., Mandrigin, A., Tajadura-Jimenex, A. & Lopez, C. (2013). Bodily ownership and self-location: Components of bodily self-consciousness. *Consciousness and Cognition*, 22 (4), 1239-1252. [10.1016/j.concog.2013.08.013](https://doi.org/10.1016/j.concog.2013.08.013)
- Shoemaker, S. (1968). Self-reference and self-awareness. *Journal of Philosophy*, 65 (19), 555-567. [10.2307/2024121](https://doi.org/10.2307/2024121)
- (1970). Persons and their pasts. *American Philosophical Quarterly*, 7 (4), 269-285.

- (1996). *The first-person perspective and other essays*. Cambridge, UK: Cambridge University Press.
- Singer, W. (2004). Synchrony, oscillations, and relational codes. In L. Chalupa & J. Werner (Eds.) *The visual neurosciences* (pp. 1665-1681). Cambridge, MA: MIT Press.
- (2007). Large scale temporal coordination of cortical activity as prerequisite for conscious experience. In M. Velmans & S. Schneider (Eds.) *Blackwell companion to consciousness* (pp. 6005-615). Malden, MA: Blackwell.
- Tsakiris, M. & Haggard, P. (2005). The rubber hand illusion revisited: Visuotactile integration and self-attribution. *Journal of Experimental Psychology: Human Perception and Performance*, 31 (1), 80-91.
[10.1037/0096-1523.31.1.80](https://doi.org/10.1037/0096-1523.31.1.80)
- Uhlhaas, P., Pipa, G., Lima, B., Melloni, L., Neuenschwander, S., Nikolic, D. & Singer, W. (2009). Neural synchrony in cortical networks: History, concept and current status. *Frontiers in Integrative Neuroscience*, 3, 1-19. [10.3389/neuro.07.017.2009](https://doi.org/10.3389/neuro.07.017.2009)
- Vallar, G. & Ronchi, R. (2009). Somatoparaphrenia: A body delusion. A review of the neuropsychological literature. *Experimental Brain Research*, 192 (3), 533-551.
[10.1007/s00221-008-1562-y](https://doi.org/10.1007/s00221-008-1562-y)
- Wittgenstein, L. (1958). *The blue and brown books*. New York, NY: Harper & Row Publishers.
- Zahavi, D. (2005). *Subjectivity and selfhood*. Cambridge, MA: MIT Press.
- Zahn, R., Talazko, J. & Ebert, D. (2008). Loss of the sense of self-ownership for perceptions of objects in a case of right inferior temporal, parieto-occipital and precentral hypometabolism. *Psychopathology*, 41 (6), 397-402. [10.1159/000158228](https://doi.org/10.1159/000158228)
- Zeng, Z., MacDonald, E., Munhall, K. & Johnsrude, I. (2011). Perceiving a stranger's voice as being one's own: A 'rubber voice' illusion? *PLoS One*, 6 (4), 1-8.
[10.1371/journal.pone.0018655](https://doi.org/10.1371/journal.pone.0018655)

Are there Counterexamples to the Immunity Principle? Some Restrictions and Clarifications

A Commentary on Caleb Liang

Oliver Haug & Marius F. Jung

Our commentary focuses on the sense of experiential ownership and its implications for the Immunity Principle. In general we think that Liang elaborates the self-as-object and the self-as-subject in an interesting and refreshing way. Nevertheless, there are some problems that we want to address. (1) First, we argue that the sense of experiential ownership cannot misrepresent the fact of experiential ownership. (2) Second, we argue that neither the sense of experiential ownership in particular nor phenomenal states in general are eligible for identity judgments. (3) Then we claim that the two alleged counterexamples actually do not provide any valid argument against IEM. (4) We close by evaluating whether it makes sense to talk about the Immunity Principle as a non-trivial property, or whether the relevant properties are just mispredication or misguided reference.

Keywords

Body-ownership | Body-swap illusion | *De re* misidentification | Fact of experiential ownership | Identification-freedom | Immunity to error through misidentification | Immunity to misguided reference | Judgments | Mispredication | Self-as-object | Self-as-subject | Sense of experiential ownership | Somatoparaphrenia | Which-object misidentification

Commentators

[Oliver Haug](#)

ruehlo1@students.uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

[Marius F. Jung](#)

mjung02@students.uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Target Author

[Caleb Liang](#)

giliang@ntu.edu.tw
國立台灣大學
National Taiwan University
Taipei, Taiwan

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Introduction: Preliminaries and conceptual clarification

Liang investigates some interesting issues concerning self-consciousness and its relation to conscious phenomenology and bodily self-con-

sciousness. His argumentation, which has the aim of being interdisciplinary fruitful, is closely tied to some conceptual distinctions that are

also very important for our commentary. First, he refines the initial point of the Wittgensteinian distinction between self-as-object and self-as-subject (Wittgenstein 1958). An important distinction concerning the former is the sense of body ownership and the sense of self as physical body, which describes the self-as-object in a more fine-grained manner. The self as subject is also sub-classified in terms of the fact of experiential ownership and the sense of experiential ownership. The sense of experiential ownership describes mental states that refine proprietarily aspects of *who* is having the experience in question. Liang claims that the sense of experiential ownership is not privileged in the sense that it gives rise to the well-known property *immunity to error through misidentification* (IEM). In the second part of his investigation he is concerned with theoretical and empirical investigations made by Damasio, Panksepp and Northoff, which do not provide substantial evidence in their measurements for the sense of self as experiential subject. They rather concern the self-as-object and therefore disregard substantial aspects of self-consciousness. Our commentary will focus on the sense of experiential ownership with regard to IEM. According to Liang, there are several counterexamples to IEM, mainly to be found in misrepresentations (like in the body-swap illusion) due to a sense of experiential ownership. In this commentary, we ask ourselves the following questions: is the sense of experiential ownership a plausible candidate for exemplifying the property of IEM, and could there be serious counterexamples to that principle? We defend the following four theses:

- (1) The sense of experiential ownership cannot misrepresent the fact of experiential ownership (cf. section 3).
- (2) Phenomenal states like the sense of self as experiential subject are ineligible to serve as *bearers* of IEM as a property (cf. section 3).
- (3) Liang's counterexamples do not provide real counterexamples to IEM, be-

cause they do not aim at the target phenomenon (cf. section 4).

- (4) IEM is either a very trivial property of judgments or beliefs or could be explained in terms of immunity to misguided reference (cf. section 5).

In order to defend these four theses we introduce two conceptual distinctions by which we hope to describe the target phenomenon in greater detail. Some philosophers, such as Evans (1982) and Shoemaker (1968) consider IEM to be a property of *judgments*, whereas others, such as Coliva (2002) and Bermúdez (1998), talk about some phenomenal aspects. Let us summarise these two accounts of IEM as follows:

First-person pronoun immunity (IEM-FP): A speaker who uses the singular indexical expression “I” knows a thing to be φ and conducts a predication “a is φ ”. This judgment is based on the rule of *identification-freedom*, so that it is clear that “I am φ ” is a judgment that does not depend on any further identification component.¹

Phenomenological immunity (IEM-P): Immunity to error through misidentification is a property of phenomenal states that characterises the constituents of first-person judgments. These identification-free constituents manifest themselves in phenomenological experiences about oneself.²

1 In other words, a judgment is identification-free if to judge that “a is φ ” *eo ipso* is to judge that “I am φ ”. Shoemaker's argument for identification-freedom (subject-use) can be summarized as follows. (1) The utterance “a is φ ” gives rise to an error through misidentification, if a speaker knows a thing to be φ and mistakenly thinks that ‘a’ refers to φ (cf. Shoemaker 1968). (2) Not every subject-use, which can give rise to knowledge about oneself, depends on identification, because this would lead to an infinite regress (cf. *ibid.*). (3) Since there is no identification of an object with a thinker in subjective first-person judgments, they are clearly incorrigible (relative to the first-person pronoun (e.g., some proprioceptive judgments or “I feel pain”; Shoemaker 1968). (4) Since the use-as-subject does not depend on identification, an error through misidentification is impossible.

2 This is a highly controversial metaphysical generalization of IEM, because it assumes that *there are* phenomenal constituents of IEM that serve for IEM as a property of judgments. Lane (2012), for instance, denies that there are any unique constituents that could explain *mineness* or mental ownership. Nonetheless, we suspect that the authors who defend theories of phenomenological immunity, like Liang, have to accept this generalization in one or another way. François Recanati (2012) seems to defend a similar position. A sub-

There is a strong inclination that the above philosophers who describe IEM as a property of judgments claim IEM to be something like a conceptual truth. But this would be overhasty, because of the fact that it is not yet clearly elaborated what a judgment with regard to the property in question actually is. We turn to this problem later. Liang seems to be a proponent of IEM-P, which holds that IEM is a property of phenomenal states:

My target is a form of mental immunity that I call experiential immunity. Experiential immunity concerns phenomenal experiences. It is a form of relative immunity—that is, it is relative to first-personal access to phenomenal states, such as introspection, somatosensation, proprioception, etc. (Liang [this collection](#), p. 8)

What distinguishes Liang’s account from others is that he emphasises that IEM does not hold necessarily. In an older paper he and Lane state that the philosophical orthodoxy of IEM has never been empirically challenged. That is because the majority of philosophers hold IEM as a conceptual truth, which has nothing to do with the empirically-tractable structure of reality (cf. [Lane & Liang 2011](#)). Our commentary is structured as follows. First, we summarize Liang’s most interesting claims and distinctions (cf. section 2). In section 3 we claim that it is impossible that the sense of experiential ownership can misrepresent the fact of experiential ownership, and that phenomenal states are not eligible bearers of IEM as a property. In section 4 our main claim is that Liang’s interpretation

subject experiences a state, for instance, through a proprioceptive mode, whereas the subject is not explicitly represented. He calls this *implicit de se* immunity to error through misidentification (IEM). This mode of experience is immune to error through misidentification because it is identification-free. Then the subject reflects upon this mode of experience, which means that she represents explicitly *who* the subject is. This is the *explicit de se*. The explicit involvement of a subject is constituted by the implicit involvement of the subject, which is identification-free. Since the former, the constituent, is IEM, it is also the latter. Recanati’s argumentation was the inspiration for summarizing proponents of phenomenal (or perceptual) immunity, as we did with IEM-P. The question arises whether some systems without any instantiated phenomenal properties could have beliefs that have the property of IEM. Since we are skeptical about IEM-P as a constituent of IEM-FP, as will be argued, nothing excludes this possibility according to our account.

of some empirical studies does not provide counterexamples to IEM. Section 5 develops the consequences of this claim and concludes with some aspects concerning the way in which we could talk about IEM in a more deflationary and less mysterious manner, such as in terms of immunity to misguided reference (IMR) or mispredication. In section 6 we conclude with some proposals for future research.

2 The sense of body ownership vs. the sense of experiential ownership

Before we discuss the self-as-subject in a more detailed manner, we focus on Liang’s conceptual refinements of the self-as-object. Liang proposes three important distinctions that are very helpful for the debate on bodily self-consciousness. The first marks out the fact of body ownership and the sense of body ownership. The fact of body ownership has nothing to do with phenomenal experiences of one’s own body. It just describes “[...] a biological fact about the anatomical structures of one’s body” (Liang [this collection](#), p. 2). In contrast, the sense of body ownership describes the experiences of the factual aspect of body ownership. Hence, to experience something as belonging to one’s own body is to experience a biological fact. Then Liang distinguishes between *the first-personal sense* and *the third-personal sense of body ownership*. We think that this is a very explanatorily fruitful distinction. The first-personal sense of body ownership describes some pre-reflective states such as walking or proprioceptive states. But these states could be third-personal or reflective as well if there are experienced from the outside, for instance through mirror recognition of one’s own body parts.

The last distinction concerning the self-as-object is between *the sense of body ownership* and *the sense of self as physical body*. The sense of body ownership is the experience of various body parts belonging to one’s own body, while the sense of self as a physical body concerns more ontological questions of the self. Here Liang introduces the sense of self as physical body as the sense of being a person of flesh and blood.

Let us concentrate on the distinction between *the first-personal sense* and the *third-personal sense of body ownership*. For us it is a rich conceptual tool that can help us refine the classic Wittgensteinian distinction between self-as-object and self-as-subject. We suggest that the notions of the first-personal sense of body ownership and the sense of experiential ownership are often used interchangeably. There are closely related but of course distinct from each other. Imagine a person who recognises that her legs are crossed through the first-personal sense of body ownership. She experiences her legs to be her own crossed legs. But here the Wittgenstein question makes perfect sense. Is it really she who is experiencing that very state? This open question marks out the sense of experiential ownership. We share Liang's criticism that the lack of a distinction between a sense of bodily ownership and a sense of experiential ownership could result in overinterpretation of some empirical data. If this distinction makes sense—as we think it does—then Liang's claim that Damasio, Panksepp, and Northoff's conceptions of the core self do not target the sense of self as experiential ownership sufficiently is plausible. The claims fit rather with *the first-personal sense of body ownership*.

In order to target the sense of experiential ownership, the Wittgenstein question could be asked to the participants of some experiments. Then we could, according to Liang, measure and elaborate on not only *what* is experienced but also on *who* is experiencing. Liang convinces us that there is more to explain than just senses of body ownership. If the sense of experiential ownership marks out a specific phenomenal target property, then much has to be done in philosophical and interdisciplinary empirical research. If Liang is right—which we think he is—and the target phenomenon of the sense of experiential ownership is empirically tractable, some further research would be very interesting and illuminating.

3 IEM-P—A conceptual matter?

In order to discuss this appropriately we first have to recall some of Liang's conceptual refine-

ments. One important distinction we want to discuss is the distinction between the fact of experiential ownership and the sense of experiential ownership, which mark out the factual and the subjective aspect of experiential ownership. The third-personal sense of experiential ownership describes the factual aspect, which can be observed from the outside via fMRI. Liang calls it a biological fact that, when a subject undergoes an experience, there is an objective fact of experiential ownership that is constitutive of the sense of experiential ownership. The first-personal sense is a phenomenal property of mental states, which means that it does not require further informational states to ensure that the one *who* is experiencing it *from the inside* sense herself experiencing it, which would be the “for-me” aspect. This is the property which concerns the aspect in which we and Liang are interested in: the self-as-subject. In order to evaluate the arguments of IEM, Liang uses the conceptual refinement offered by Pryor (1999), namely the *de re* and *which-object misidentification*. The former has been challenged through cases of somatoparaphrenia, the latter by the so-called body-swap illusion, both of which provide cases of misrepresentation. What happens in cases of misrepresentation? For Liang the sense of experiential ownership misrepresents the fact of experiential ownership. We argue that there are some aspects of the fact of experiential ownership and the sense of experiential ownership that are not that clear. Our thesis is that the fact of experiential ownership has nothing to do with IEM-P in the first place, but is rather what some philosophers describe as the conceptual truth of a subject having an experience. If you are describing the specific phenomenological richness of an instantiated experience, it is obviously true that it is an experience of a subject.³ Since subjects are the bearers of experiences (as opposed to objects) it is quite obvious that there is a *fact* that somebody

³ It is important to mention that a subject can experience a state “from the inside”, which she does not experience as her own. Experienced “from the inside”, it could belong to someone else or to nobody (Lane 2012). But this fact, which we take to be an analytic truth, is something ascribed “from the outside”. To say that somebody has an experience, is just to say that the experience is instantiated in a subject, regardless of which experience the subject undergoes exactly.

has this experience. This can be illustrated in Liang's own words: "[w]hen a subject experiences a phenomenal state, there exists a fact that he is the subject of that state" (Liang [this collection](#), p. 6). But this is just analytically true, since experiences are not free-floating occurrences—because they, as a matter of principle, have a subject of experience. This is about using the words "somebody's experience" correctly and is rather a description from the outside. It tells us nothing substantial about IEM-P. Perry (1998, pp. 96–97) talks about a similar phenomenon while recapitulating Locke's idea of personal identity. He claims that "[a]n instance of being aware of an experience, and the experience of which one is aware is known, necessarily belong to the same person [...]". To say something substantial it would be important for the content of the phenomenal experience of a specific state to *concern* the subject itself. But the content, experienced "from the inside", is of course different from an analytical truth, because phenomenal states have nothing to do with the right usage of words. The content of the phenomenal experience is what Liang calls the sense of experiential ownership, experienced *from the inside*. Granted that these two conceptualizations are correct, it is impossible that a phenomenal state like the experiential ownership represented from the inside can misrepresent something that is rather a conceptual ascription or description from the outside. They are completely different categories. To understand this we can think of a patient suffering from dissociative identity disorder (DID), who has many different personalities. What would be the fact of experiential ownership here? To answer this question a very specific and rigorous conception of personal identity is needed, which cannot be discussed here.

Let us summarise the argument:

Sense of experiential ownership cannot misrepresent fact of experiential ownership

(1) The fact of experiential ownership is to describe (as we see it), as a matter of logical necessity, that an experience is instantiated in a subject, that is (according to Liang), if a sub-

ject undergoes an experience in the actual world, a matter of fact.

(2) The sense of experiential ownership concerns the content of a phenomenal experience, which can either be experienced as owned by a subject or by nobody.

(3) Phenomenal experiences do not represent facts or states of affairs and even less analytic truths.

(4) The sense of experiential ownership cannot represent the fact of experiential ownership.

(5) A representation necessarily goes together with the possibility of a misrepresentation.

(C) The sense of experiential ownership cannot *misrepresent* the fact of experiential ownership.

Does it generally make sense to talk about IEM-P as a property of phenomenal states? The remaining story about IEM-P could be that it serves as the basis for judgments that usher in beliefs (see section 4). The immunity would then hold just through the structure of experience itself. But does it?

We claim that there no error through misidentification is possible, because of the lack of judgments and cognitive elaboration at the phenomenal level. An identity judgment requires identifying two conceptually-represented ingredients. Phenomenal states can be accompanied by conceptual ingredients, but they are not basic properties of phenomenal states themselves.⁴ Thus, they are distinct from one another. Hence we could say that phenomenal states are neither *eligible* for such a kind of error in general nor for a *de re* or which-object misidentification in particular. The intelligibility of IEM-P is very doubtful. Let us again summarise the argument:

Ineligibility of IEM-P

(1) To talk about identification is to talk about judgments and inferences that can be identified with one another, which means that they are *judged* to be identical.

⁴ Proponents of *Cognitive Phenomenology* would probably deny this claim. We stick with Carruthers & Veillet (2011), who says that cognitive thoughts could causally initiate some phenomenal experiences. The stronger claim, that thoughts constitute phenomenal experiences, lacks substantial argument. Hence we stick to the position that phenomenal states and thought contents could occur in isolation from each other.

(2) To talk about misidentification is to talk about some defective judgments.

(3) Phenomenal states have nothing to do with judgments and inferences in the first place.

(C) Phenomenal states lack the basic properties to be defective.

The ineligibility of phenomenal states of course satisfies the rule of identification-freedom. But since phenomenal states are always identification-free, the claim that they are immune to error through misidentification is misleading. Why is that? Remember that the content of phenomenal experience could occur without being owned by somebody (Lane 2012). Nevertheless, an experience is instantiated in a subject, which is just a matter of principle or the factual aspect. If the content of a phenomenal experience just occurs, without an experience of *mineness*, then the rule of identification-freedom tells us nothing substantial, because of the lack of any committed judgment. An interesting question, of course, is whether there are any judgments that are identification-free.

We would recommend talking about IEM as a property of judgments or beliefs (IEM-FP) instead of talking about phenomenal states. Nevertheless, there are also some problems with IEM-FP that we will present and discuss in section 4. Let us now have a closer look at the two alleged counterexamples that Liang proposes.

4 Two counterexamples to IEM-FP?

IEM is generally considered to be a property of judgments concerning the first-person perspective and respectively involving the first-person pronoun. A major problem in the current discussion about IEM is that no solid account of what judgments are is given. In contrast to philosophers that are concerned with beliefs, who usually give a brief declaration of what they take beliefs to be (e.g., relations, sentence operations etc.), philosophers involved in the IEM discussion seem to take judgments to be already widely understood. Since the initial paper written by Shoemaker (1968) focuses on the identification-freedom of judgments, we think that what philosophers

usually talk about using the term “judgment” is inference or reasoning.⁵

So we take judgments to consist of propositional reasoning. Let us have a look at some examples:

Judgment A:

- (1) John is a fish. (Fa)
- (2) Fish can swim. $(\forall x)(Fx \rightarrow Gx)$
- (C) John can swim. (Ga)

Judgment B:

- (1) John is a fish. (Fa)
- (2) John is Jim’s best friend. (a=b)
- (C) Jim’s best friend is a fish. (Fb)

Though usually the conclusion of these inferences is what is referred to using the term “judgment”, we do not think that philosophers generally tend to take judgments as being adequately analysed as propositional attitudes (as relations between persons and propositions like

⁵ The reason for this is the following: if you talk about “identification-components”, there must be something that is composed of at least one identification-component, and probably of something else as well. The identification component (as described by current philosophers— $a=b$) is either a sentence or a proposition, either expressing an identification or representing it. (This distinction is just made to satisfy Platonists and nominalists.) What is it that is composed of identification- and other components? We think, according to the usual use of language of philosophers debating IEM (She sees a bleeding hand in the mirror and thus judges “I am bleeding”), that the most probable answer is that they are part of an inference. Whenever you say that one “judges” p , you want to express not only that she believes p , but also that she has come to this belief through inference. We take this to be an adequate interpretation of the term “judgment” as used in Shoemaker, Pryor, Barz, and probably Liang as well. It is probably inadequate for every instance of “judgment” in philosophy, because our interpretation suggests that there are (hidden or opaque) processes that are important for calling something a judgment. Even though proponents of accounts that are Rylean (Ryle 2009), for example, would strongly disagree (because they wouldn’t accept that there are hidden processes that we want to talk about using the term “judgment”), we think that there is in fact an ontological or categorical difference between judgment and beliefs: either judgments *are* processes and beliefs *are* states, or judgments are a subclass of beliefs, but a subclass of beliefs that one has come to through a process of inference (which is not necessarily the case with beliefs—just imagine someone manipulating your brain such that you gain new beliefs). So, unlike Ryle, we would say that as long as we are talking about human beings, judgments are certainly something that happen in the hidden depths of the human brain. And we can represent them—for our purposes—as structured like logical inferences. Please note that this is just an additional remark concerning our positive account of judgments that a lot of papers seem to lack. Our central argumentation does not rely on this specific ontological reading of “judgment” and “belief”.

“Jim believes that it is raining”). We take the whole inference to be what is referred to with the term “judgment”, and the conclusion to be what is referred to using the term “belief”.⁶ Judgments A and B are analogous in the following sense: they are both judgments involving two premises and their logically necessary conclusions. But they differ in a particular aspect that is of the highest importance concerning IEM: only judgment B involves an identification, whereas judgment A is identification-free. So the first thing we can say is that IEM following from identification-freedom is not an exclusive property of judgments involving the first-person pronoun—there are numerous judgments that do not contain any identification-components. This is our first reason for thinking that IEM may hold, but is not a remarkable or significant property exclusively reserved for judgments involving the first-person pronoun.

What Shoemaker wants to make clear is that there are certain judgments that cannot take the logical form of judgment B and that these judgments involve the first-person pronoun, in the sense of Wittgenstein’s “subject-use”. Let us take a look at what Shoemaker means by giving examples for the object-use and the subject-use:

Object-use:

- (1) The person in the mirror is looking tired. (Fa)
- (2) I am the person in the mirror. (a=b)

- (C) I am looking tired. (Fb)

As you can see, there are judgments involving the first-person pronoun that also involve an identification-component—at least that is what Shoemaker (1968) thinks. But when he claims that there are judgments that are immune to error through misidentification, he does not claim that they are immune to *any* error, and nor does he claim that the identity relation holds with metaphysical necessity—he just

claims that whenever one judges and this judgment involves certain kind of predicates (or properties) it is automatically identification-free. One of those predicates (or properties) is being in pain. Let’s have a look at how the judgment would work with this special predicate that we may call P*.

Subject-use:

- (1) There is something that is in pain. ($\exists x$) (P*x)
- (2) P* is always a property of the person recognizing it. (P*gen)

- (C) I am in pain (P*a)

In fact this formal representation of such a judgment is even weaker than what Shoemaker may have had in mind, thus the strong reading of his idea of judgments that are IEM because of their identification freedom would be:

- (1) There is something that is in pain. ($\exists x$) (P*x)

- (C) I am in pain. (P*a)

This reading gets closer to Shoemaker’s idea, because he would not agree that judgments explicitly involve a generalization such as (2). We undertook this brief exercise first of all to put pressure on the following point: although philosophers of different generations have been talking about IEM for decades, they usually fail to give an *explicit account* of what judgments are and how they work.⁷ This exercise was meant to fill this theoretical gap for the purpose of the current discussion. So whenever someone utters “John is a fisherman”, we take this sentence to express a propositional attitude—a belief. But when we believe that he *judges* “John is a fisherman”, we also take that person to have made an inference, simply because that is what we want to say when we ascribe a judgment to him. Shoemaker’s claim that judgments like “I am in pain” are immune to error through misidentification does not mean that there are hidden structures, neither of the sentence ex-

⁶ Note that we in fact think that beliefs are brain states and judgments (if they are inferences) are cognitive processes—but they do not need to be brain states and cognitive processes. Depending on which understanding of propositions you prefer (e.g., the meaning of sentences or informational packs), any machine that is capable of some kind of reasoning can judge and have beliefs.

⁷ This means that there are no papers about IEM that give a positive account of judgments, e.g., Shoemaker (1968), Evans (1982), Barz (2010).

pressing the judgment nor of the propositional attitude expressed by the sentence “I am in pain”; it means that no identification-component was involved in the inference that has been made.

The second reason for undertaking this exercise is that we want to have a look at whether Liang’s counterexamples (especially the somatoparaphrenia example) are real counterexamples. We do not think that the two examples Liang gives are in any way counterexamples to IEM—though they are philosophically very interesting, especially concerning theories of self-consciousness. Liang claims that the two counterexamples falsify the Immunity Principle, but we claim that they do not meet the conditions that have to be met to falsify this theory. So we must first see what Liang takes to falsify the IEM theory and then settle on a criterion for how the IEM theory could be falsified.

Liang thinks that the following would suffice for IEM to hold:

- (1) for every phenomenal state there must be a subject who experiences it; (2) every phenomenal state is in principle available to first-personal access (Shoemaker 1996); (3) every phenomenal state is experienced by the one who has first-personal access to that state. The crucial point is that (1)–(3) do not imply that (4) every phenomenal state is, from the first-person point of view, represented as experienced by the one who has first-personal access to that state. (Liang this collection, p. 8)

Liang also considers his two counterexamples (the somatoparaphrenia patient and the body-swap illusion) to be counterexamples to (4), so the IEM-principle does not hold. In fact we agree with Liang that at least one of these examples is a counterexample to (4) but we do not agree that (4) is necessary for IEM to hold. So let us first have a look at how a falsification of the IEM-theory would have to look. The IEM-theory comes in the form of a material conditional: *if* a person judges “I am φ ”, *then* she cannot be wrong because of a misidentification. The truth conditions for a material condi-

tional are clear: the conditional is wrong if and only if the antecedent is true and the consequent is wrong. This brings us to the definition of a theoretical falsification of IEM:

Falsification of IEM =_{Df} : **1.** The IEM-theory would be falsified if and only if a person judges “I am φ ” and is wrong in her judgment because of a misidentification. Or, more precisely: **2.** The IEM-theory would be falsified if and only if there is an example of a person that believes “I am φ ” and comes to this belief through inference (the judgment) that involves an identification component and this identification is wrong.

Thus, speaking more formally, a judgment of the following two forms must be present (cf. Pryor 1999):

wh-judgment⁸

- (1) $(\exists x)(Fx)$ (predication to a variable)
 - (2) I am x (identification, $x=a$)
-
- (C) I am F (predication to a constant, depending on the identification), (Fa)

or

de re judgment⁹

- (1) A particular thing (*de re*) is F (Fa)
 - (2) I am that particular thing ($a=b$)
-
- (C) I am F (Fb)

We pick these two different structures to emphasise that Shoemaker did not exclusively talk about *de re* “attitudes” but also about “existential quantification”, though he did not do so explicitly. Note that besides the presence of belief states such as (c) it is necessary for the falsification of IEM-theory that this conclusion is only wrong because (2) is wrong.

⁸ A *wh*-judgment involving an identification starts with existential quantification over a variable. You know that there is *something* that has a particular property and then you identify that something with, e.g., yourself.

⁹ A *de re* judgment involving an identification starts with a predication to a particular thing—a constant. So you know a *particular* thing to have a particular property and then you find that thing to be identical with, e.g., yourself.

The crucial question concerning Liang's counterexamples is: do they meet this condition? Consider the first example. It is—as Liang sees it according to Pryor—an example of a *de re* misidentification. *De re* misidentifications occur, for example, when there are two objects equally eligible for exemplifying the property in question. To show that Liang's first example is a counterexample to IEM one would have to prove, first, that the structure of a *de re* judgment as stated above holds, and second that the judgment is only wrong because the second premise is not true. Recall the first experiment: a patient suffering from somatoparaphrenia, FB, is touched on her hand and asked whether she feels her hand being touched. She answers “No”. When she is asked whether she feels her niece's hand being touched, she gives a positive answer (FB believes that her hand is in fact her niece's hand, and has been placed on her body). But since she does not judge “I am being touched on my hand”, the necessary conditions for falsifying the IEM-theory are not met. The material conditional could only be proved wrong if the antecedent (a person judging that she has a certain property) is true, but in this case it is not true. The conditions would have been met if she had answered “I feel being touched on my hand”, even though she was not, and even though the only reason why she was wrong was because she misidentified her own sensations with someone else's. But she does not commit the error of judging “I am being touched” in the first place, so the IEM-theory is not falsified. It is crucial here to understand that falsification of IEM does not depend on what exactly she said, but whether she judged that she had a certain property. Unfortunately wrongly judging that one is not touched, though one is touched, does not get close to a falsification of IEM, by definition of the truth conditions of material conditionals.

Now let us have a look at Liang's second counterexample: the body-swap illusion. This, according to Liang and Pryor, is an example of a *wh*-misidentification that happens when someone simply knows a property to be there (e.g., a smell) and falsely ascribes this property to a particular object. In this setup, the participants

judge that they are shaking hands with themselves. This example gets much closer to the claim of IEM, because they in fact judge, and judge falsely, that they experience something, and there *is* another person who really seems to have that experience. So it seems that one of the following inferences is made:

- | |
|--|
| (1) A particular person is shaking hands with myself. (Fa) |
| (2) I am that person. (a=b) |
| (C) I am shaking hands with myself. (Fb) |

or

- | |
|---|
| (1) There is something that is shaking hands with myself. ($\exists x$)(Fx) |
| (2) I am that something. (x=a) |
| (C) I am shaking hands with myself. (Fa) |

If these judgments occurred it is obvious that they are false because the second premise is false—thus an error through misidentification was made. But did the participants really commit such an error? Recall that the IEM thesis would be falsified if a person believed a certain proposition but was mistaken because and only because she misidentified herself with someone else. Did the participants really believe that they were shaking hands with themselves? We assume that they most certainly did not. Of course they remarked that they were shaking hands with themselves, but we take them to speak merely metaphorically and not literally. If one wanted to be sure, the same experiment would have to be made, asking the participants whether they believed that they were shaking hand with themselves (not if it felt as if they were). Even if they believed that they were shaking hands with themselves, the judgment would probably not have the form stated above, because they did not have the experience of the other participant—only if they had an experience that that very (exactly the same) experience depended upon another person, and only if they accidentally identified themselves with that person—only in that case would the IEM-theory be falsified. But the participants did not have the experience of the other person wearing the

camera. They were having their very own experience—caused by the informational flow starting with the display (monitoring not the perspective of the person wearing the camera, but the camera’s perspective) and their own lenses, their own retina, and so on. The experience they ascribed to themselves was not the experience of another person or agent, it was their own experience. They were only wrong in judging that they were shaking hands with themselves because they in fact did not shake hands with themselves. This is not a misidentification but simply a mispredication. This problem will be elaborated in section 5.

So why does Liang think that these examples are counterexamples to IEM? Because he takes (4) to be crucial for IEM to hold. The differences between what Shoemaker and Evans take to be the theory of IEM and what Liang takes it to be are the following:

Shoemaker/Evans: If a person believes that she has certain properties, she cannot be mistaken in having them by misidentifying herself (or her phenomenal states) with someone else or someone else’s states. In this conditional, the antecedent implies a person to believe something about herself or, speaking in Liang’s terms, a person to represent herself as having a so-and-so experience. But Liang’s conditional looks quite different:

Liang: “(4) every phenomenal state is, from the first-person point of view, *represented as* experienced by the one who has first-personal access to that state.” (Liang [this collection](#), p. 8)

So what used to be the antecedent in the original theory becomes the consequent in Liang’s theory—thus Liang is right that (4) does not hold and that the somatoparaphrenia patient and her reports are counterexamples to (4), but he is not right in taking this fact to falsify the IEM-theory.

5 Why does IEM-FP hold?

There seems to be an immunity relative to the first-person pronoun, which at least guarantees that you cannot have a belief like “I believe that I am in pain” and accidentally take

someone else to have that belief. It probably also guarantees that in this case you cannot be wrong about who is in pain. We think that there are a few good theoretical candidates for explaining this kind of immunity. These candidates are:

1. Irrelevance of misidentification
2. Immunity to misguided reference
3. Reference magnetism

Since reference magnetism¹⁰ is a highly controversial, metaphysical notion and it would take too much time to elaborate this view correctly (which would certainly include a refreshment of Lewis’ philosophy of reference), we will focus on the first two for the sake of this commentary.

1. Irrelevance of misidentification:

If you take judgments about yourself to be a) always starting with *de re* beliefs and b) single-predicative in form, it seems impossible to construe misidentification as being relevant to the truth-value of a sentence or proposition. This point has been made by Barz (2010). Barz takes the current discussion to assume that there are two fundamentally different kinds of errors that can occur: an error through misidentification and an error through mispredication. It should be clear what an error through mispredication is supposed to be: an error through mispredication occurs when a person’s judgment is wrong and is only wrong because the predicate she thinks applies to a particular object in fact does not apply to that object. Barz’ definition of an error through misidentification (in general) is the following:

General error through misidentification (EM-G): A person S (i) believes (*de re*) of a certain thing that it is *F*, (ii) believes that

¹⁰ A short explanation: reference magnetism is a theory that claims that there are metaphysically distinguished objects of reference in the world (no matter whether they are abstract or concrete) that function as magnets for certain expressions. This could, for example, hold for natural kinds and existential quantification. In the case of existential quantification, some philosophers, like Theodore Sider (2009), claim that there is no possibility of talking about *existence* without talking about the very same thing that everybody talks about—as long as there is no explicit or implicit quantifier restriction. Reference Magnetism plays an important role in the debate about quantifier variance and verbal debates.

thing to be identical with a , and (iii) thus judges that a is F . But (iv) a is not identical with the thing S believes to be F .

According to Barz this kind of error cannot happen at all, so the proponents of the IEM-theory are right—but in fact IEM is not an exclusive property of judgments concerning the first-person or involving the first-person pronoun, and is instead a property of any judgment. His argumentation can be summarised in one sentence: since there are examples of judgments involving misidentification that are nevertheless true, and since there cannot be judgments involving mispredication that are true, there are no errors through misidentification. A judgment is right or wrong solely depending on whether the predicate applies to the object.

Imagine the following situation that is usually used to distinguish between notional and referential use of singular terms: Peter is a detective, investigating the case of Smith's murder. Participating in the judicial proceedings, a man, accused of having murdered Smith, behaves so strangely that Peter, the detective, judges: Smith's murderer is a maniac. He is using the term "Smith's murderer" to refer to the person that is accused of having murdered Smith, and according to most theories of reference he does in fact refer to that person with that term. But what if that person is not the one who murdered Smith, but is nevertheless still a maniac? Thus a misidentification has occurred, but no error. On the other hand, if the person were Smith's murderer but not a maniac (maybe his weird behaviour was the result of pharmaceutical treatment)—Peter's judgment would be wrong.

The same goes for the traditional wrestler example. Imagine that wrestler A and wrestler B are in a close wrestling fight and wrestler A does not misidentify her arm with the arm of wrestler B but still, for some strange reason—maybe there are blood smears caused by a bleeding bird that flew over the two wrestlers—comes to judge "My arm is bleeding" (although wrestler B's arm is actually bleeding). She would be wrong, but her error would not be one of misidentification but of mispredication. Thus, as Barz believes, there are no errors through misidentification, because the

only thing that necessarily suffices for the falsity of a judgment is mispredication.

As one can guess, Barz' theory does not completely fit with our theory of judgments. While we take judgments to be processes of inference, thus involving several propositions, Barz seems to take judgments to be relations to single, structured propositions. We can agree with Barz if he can explain how the identification component in the judgment—which would, in our terms, be one of the premises used during the inference—is in fact a kind of predication.

2. Immunity to misguided reference:

Howell (2007) wants to distinguish between two kinds of immunity: immunity to error through misidentification and immunity to misguided reference:

IEM is often confused with what I call Immunity to Misguided Reference (IMR). A judgment that x is F has IMR if it is impossible for someone to make that judgment while being mistaken about the reference of x . All I-judgments have IMR, while not all I-judgments are IEM. (Howell 2007, p. 584)

To say that there is something like immunity to misguided reference (IMR) does not mean that one can never be wrong about the reference of any term one uses. It just means that whenever you want to refer to yourself using the term "I" you cannot fail to do so.

We think that a majority of the proponents of IEM are in fact proponents of IMR. And because IEM is thought to be an immunity relative to the first-person pronoun (what we have termed IEM-FP), it makes sense to say that this immunity is in fact an immunity of referring acts in general and not of judgments exclusively. Talking about IMR can be helpful in two ways: first, it can be helpful in stressing the fact that IEM is not a theory about the self or about subjectivity but simply a theory about linguistic rules and reference. Thus IEM-FP is a trivial property that can be explained by the semantic rules of usage of the word "I".

Second, it can be helpful for explaining our intuitions in complicated cases of self-refer-

ence and by determining the objects of beliefs. Think of the two wrestlers again. When one of the wrestlers states “I am bleeding” or “My arm is bleeding”, she is wrong, but it seems as if she is not necessarily wrong because of a misidentification. Let’s have a look:

- (1) Wrestler A correctly describes her belief, intending to refer to herself using the first-person pronoun.
- (2) One cannot fail to refer to oneself when using the first-person pronoun. (IMR-rule)
- (3) Wrestler A has a belief about herself (granted by accepting 1 and 2).

So far the argument is trivial—stating that Wrestler A has a belief about herself just means that she has *any* kind of belief. It does not show that Wrestler A has a *de re* belief about herself. This comes from the second part of the argument:

- (4) A *de re* belief is a belief that holds if the believer is in a non-conceptual, contextual relation to the object the belief is about.¹¹
- (5) One is always in a non-conceptual, contextual relation to oneself.¹²
- (C) Wrestler A has a *de re* belief about herself (granted by accepting 3, 4 and 5).

Opponents of the IEM-theory would have to state that wrestler A has no *de re* belief about herself, because the object her belief is really about is not herself, but wrestler B, misidentified with herself (thus creating a *de dicto* belief about herself and a *de re* belief about wrestler B). But by accepting IMR and certain accounts of *de re* attitudes we can see that wrestler A’s attitude is a possible candidate for a *de re* belief about herself. Thus the only reason why she would be wrong is—as we have seen above—mispredication.

6 Concluding remarks

The question with which we began was how the sense of experiential ownership is related to the well-known property of IEM, and whether, if it

is, the proposed counterexamples are cogent. First of all we argued that it is impossible to talk about the sense of experiential ownership misrepresenting the fact of experiential ownership, since the latter is a conceptual ascription from the outside that has nothing to do with phenomenal states that are experienced from the inside (cf. thesis 1). Second, IEM-P is an incoherent notion, because phenomenal states lack the basic properties that are possessed by judgments and inferences, namely to be defective—which suffices for a misidentification. Since they lack these properties, the claim that phenomenal states are immune to error through misidentification is misleading (cf. thesis 2). Third, we argued that the alleged counterexamples to IEM are just counterexamples of Liang’s fourth premise. But premise four is not necessary for IEM to hold. In any case, the counterexamples do not seriously challenge IEM, because the necessary conditions for a falsification are not met (cf. thesis 3). The last section addressed some aspects concerning how to talk about IEM convincingly in future philosophical research. Our suggestion is somehow deflationary, since it is not necessary, but very likely that the more interesting properties for talking about are mispredication and IMR (cf. thesis 4).

We are looking forward to the time when philosophical as well as empirical interdisciplinary research concerning the mind focuses on Liang’s commitments on self-consciousness, most interestingly the sense of experiential ownership. We think that this *explanandum* has not yet been enriched with empirical data. Here Liang perhaps provides a good starting point for future research. In order to provide fruitful data, we think that to ask the Wittgenstein question, as Liang proposes, is a promising idea. But nonetheless the question has to be subdivided in order to provide a fruitful questionnaire. Here are some proposals that are, of course, provisory, which could be more fine-grained, depending on the experiment:

On a scale from 1 to 10, how much do you feel the experience as being owned by you? Have you felt parts of your body as detached from yourself? If yes, how much were you able to control the belongingness of this body-experi-

¹¹ This is to accept a *de re/de dicto* distinction that is compatible with non-propositional attitudes.

¹² This does not mean that one is always only and exclusively in a non-conceptual relation to oneself. Of course one can have *de dicto* beliefs about oneself.

ence? Have you felt some experiences belonging to another subject, not being owned by yourself?

Here are some further theoretical questions: How is the sense of experiential ownership connected to beliefs? Could it serve to justify some beliefs? How is the sense of experiential ownership generally related to self-knowledge? We are looking forward to a fruitful discussion in philosophy of mind and in cognitive sciences with regard to the elaborated topics.

Acknowledgements

First of all, we are appreciative for the illuminating target paper. In addition to that, we would like to thank the two anonymous reviewers and Thomas Metzinger and Jennifer M. Windt for their editorial reviews. We are grateful to Thomas Metzinger and Jennifer M. Windt for the opportunity to contribute to this project. Furthermore, we want to thank Ralf Busse for the insightful suggestions with regard to some issues within IEM-accounts.

References

- Barz, W. (2010). Irrtum durch Fehlidentifikation. *Conceptus*, 93, 7-15.
- Bermúdez, J. L. (1998). *The paradox of self-consciousness*. Cambridge, MA: MIT Press.
- Carruthers, P. & Veillet, B. (2011). The case against cognitive phenomenology. In T. Bayne & M. Montague (Eds.) *Cognitive phenomenology* (pp. 35-56). Oxford, UK: Oxford University Press.
- Coliva, A. (2002). Thought insertion and immunity to error through misidentification. *Philosophy, Psychiatry, and Psychology*, 9 (1), 27-34. [10.1353/ppp.2003.0004](https://doi.org/10.1353/ppp.2003.0004)
- Evans, G. (1982). *The varieties of reference*. Oxford, UK: Oxford University Press.
- Howell, R. J. (2007). Immunity to error and subjectivity. *Canadian Journal of Philosophy*, 37 (4), 581-604. [10.1353/cjp.2008.0004](https://doi.org/10.1353/cjp.2008.0004)
- Lane, T. (2012). Toward an explanatory framework for mental ownership. *Phenomenology and the Cognitive Sciences*, 11 (2), 251-286. [10.1007/s11097-012-9252-4](https://doi.org/10.1007/s11097-012-9252-4)
- Lane, T. & Liang, C. (2011). Self-consciousness and immunity. *The Journal of Philosophy*, 108 (2), 78-99.
- Liang, C. (2015). Self-as-subject and experiential ownership. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Perry, J. (1998). Myself and "I". In M. Stamm (Ed.) *Philosophie in synthetischer Absicht* (pp. 83-103). Stuttgart, GER: Klett-Cotta.
- Pryor, J. (1999). Immunity to error through misidentification. *Philosophical Topics*, 26 (1-2), 271-304. [10.5840/philtopics1999261/246](https://doi.org/10.5840/philtopics1999261/246)
- Recanati, F. (2012). Immunity to error through misidentification: What it is and where it comes from. In S. Prosser & F. Recanati (Eds.) *Immunity to error through misidentification: New essays* (pp. 180-201). Cambridge, UK: Cambridge University Press.
- Ryle, G. (2009). *The concept of mind*. New York, NY: Routledge.
- Shoemaker, S. (1968). Self-reference and self-awareness. *The Journal of Philosophy*, 65 (19), 555-567.
- (1996). *The first-person perspective and other essays*. Cambridge, UK: Cambridge University Press.
- Sider, T. (2009). Ontological realism. In D. Chalmers, D. Manley & R. Wasserman (Eds.) *Metametaphysics. New essays in the foundation of ontology* (pp. 384-423). Oxford, UK: Oxford University Press.
- Wittgenstein, L. (1958). *The blue and the brown books (preliminary studies for the 'philosophical investigations')*. New York, NY: Harper Perennial.

Can Experiential Ownership Violate the Immunity Principle?

A Reply to Oliver Haug & Marius F. Jung

Caleb Liang

In what follows, I respond to Haug and Jung's criticisms of my target paper and defend the following claims: (1) the sense of experiential ownership can misrepresent the fact of experiential ownership; (2) the sense of experiential ownership is eligible to serve as a bearer of IEM; (3) at least some versions of IEM face genuine counterexamples; and (4) as far as the sense of self-as-subject is concerned, IEM is not a trivial property. Finally, I describe a new set of experiments that induced what I call "the self-touching illusion." The data, I suggest, strengthen the view that both the sense of self-as-subject and IEM are open to empirical as well as philosophical investigation.

Keywords

Experiential ownership | Immunity principle | Self-as-subject | Self-touching illusion

Author

[Caleb Liang](#)

yiliang@ntu.edu.tw

國立台灣大學

National Taiwan University
Taipei, Taiwan

Commentator

[Marius F. Jung](#)

mjung02@students.uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

Does the sense of self-as-subject conform to the immunity principle (IEM)? When I experience a phenomenal state, does it guarantee that based on first-personal access I cannot be wrong about whether it is me who experiences it? In "Self-as-Subject and Experiential Ownership", I elucidated the sense of self-as-subject in terms of the sense of experiential ownership, and argued that the sense of experiential ownership does not enjoy IEM. Haug and Jung raise very

substantial issues against my overall position.¹ Here, I respond to Haug and Jung's criticisms and intend to show how an interdisciplinary approach may enhance our understanding of the sense of self-as-subject.

Let me begin by suggesting that the following two issues regarding IEM are different:

¹ I am very thankful for Haug and Jung's criticisms, from which I have learnt a great deal. Below I will use "the sense of self-as-subject" and "the sense of experiential ownership" interchangeably.

(1) Does IEM correctly specify how we use the first-person pronoun “I”? (2) Does IEM really mark the line between the sense of self-as-object and the sense of self-as-subject? While (1) concerns a linguistic rule, (2) is about the nature of self-consciousness. The issue addressed in my paper was (2). I investigated the best way to understand the distinction between the sense of self-as-object and the sense of self-as-subject. I argued that IEM, or at least some versions of it, fails to draw the distinction between the two types of self-consciousness. I proposed an alternative account, according to which the distinction can be better articulated in terms of the sense of body ownership and the sense of experiential ownership.

2 Experiential ownership and the immunity principle

The first issue raised by Haug and Jung concerns whether the sense of experiential ownership could misrepresent the fact of experiential ownership at all. For ease of discussion, I will present my argument against IEM again, and then reply to Haug and Jung’s objection. Here is the argument:

- (1) For every phenomenal state there must be a subject who experiences it.
- (2) Every phenomenal state is in principle available to first-personal access.
- (3) Every phenomenal state is experienced by the one who has first-personal access to that state.

However, (1)~(3) do not imply:

- (4) Every phenomenal state is, from the first-person point of view, *represented as* experienced by the one who has first-personal access to that state (Liang [this collection](#), p. 8).

Three remarks are in order: first, when Haug and Jung characterize the fact of experiential ownership as a conceptual truth or a matter of logical necessity, what they say can be accommodated by (1) above. I agree with (1), but that is not my notion of the fact of experiential ownership. For me, the fact of experiential ownership is an *empirical* fact: it is not just that every phenomenal state has a subject; rather, it concerns exactly who is the subject of a specific

experience in a given situation. For example, right now, it is me, not you, who is experiencing back pains. So, the fact of experiential ownership is captured and fixed not by (1) but by (3) in my argument above; i.e., the question “who is the subject of that particular phenomenal state?” can be answered by ascertaining which particular subject has first-personal access to that state. Second, I would not characterize the sense of experiential ownership as concerning “the content of a phenomenal state” (Haug & Jung [this collection](#), p. 5). As I stated in the target paper (Liang [this collection](#), pp. 6–7), the representational content and the phenomenal character of a phenomenal state belong to the *what*-component of that state. The sense of experiential ownership is exclusively about the *who*-component, which is captured by (4) in my argument. Third, central to my argument is that (3) and (4) are not equivalent: as in FB’s case of somatoparaphrenia, feeling sensations is one thing, but whether she experiences herself *as* the subject of those sensations could be another. Misrepresentation may occur in one’s sense of self-as-subject when there is a mismatch between (3) and (4), i.e., when the sense of experiential ownership fails to pick out the same subject as the one settled by (3). As I suggested, the best way to describe FB’s case is that, while the fact of her experiential ownership is intact, her sense of experiential ownership fails to represent that fact. Given these remarks, the first four premises of Haug & Jung’s argument (on p. 5 of their commentary) seem to be problematic.

The second issue is about whether the sense of experiential ownership, as a phenomenal state, is eligible to serve as a bearer of IEM.² Haug and Jung insist that self-ascriptions relevant to IEM must be an explicit judgment (or belief) in an inference. However, it is not obvious that this restriction is mandatory. Given that my focus is on how to understand the sense of self-as-subject, I think that what is crucial for IEM is that the self-

² Note that, as I suggested in the target paper (Liang [this collection](#), p. 6), the fact of experiential ownership and the sense of experiential ownership are not numerically different states or events that can be detached from a phenomenal state. Rather, they are two ways of characterizing the *who*-component of that state.

ascriptions are justified on first-personal grounds, e.g., introspection, somatosensation, proprioception, etc. (cf. footnote 19 of the target paper). As the examiners of FB said: “The patient was blindfolded and instructed to say ‘yes’ when she felt a touch and ‘no’ when she did not feel any touch” (Bottini et al. 2002, p. 251). When FB said “yes” based on her sense of experiential ownership, there is no reason why this response shouldn’t count as a self-ascription. If we wish, we can reconstruct FB’s response in propositional form: I am mistaken in reporting “yes” during the test (ii) because, although I do know of someone that feels the sensations (via first-personal access), I am mistaken in thinking about who that person is. This seems to be a clear threat to IEM.

Also, it is worth pointing out that not all defenders of IEM think that self-ascriptions must explicitly be in propositional form. According to what may be called the Pre-reflective Account (Legrand 2006, 2007, 2010; Gallagher 2012; Zahavi 2005), at the pre-reflective level, the sense of self-as-subject is a constitutive component of the conscious state rather than an intentional object of consciousness. This phenomenological structure makes the sense of self-as-subject identification-free and hence enjoys IEM: when I am pre-reflectively conscious of myself-as-subject, I *cannot* be wrong about whether I am the subject of experiences. For the proponent of this account, making judgments about one’s sense of self-as-subject would count as *reflective* rather than pre-reflective self-consciousness, and hence ceases to be identification-free (Gallagher 2012, pp. 207–209). Given these considerations, I believe that the premises of Haug and Jung’s argument for the ineligibility of IEM-P are not as firm as they might think.³

The third issue is whether the specific case of somatoparaphrenia and the body swap illusion that I discussed are genuine counterexamples to IEM. The way that Haug and Jung oppose my counterexamples is related to our dispute above concerning whether IEM

has to be in the form of judgment. Haug and Jung define “judgment” as referring to a whole inference and “belief” as the conclusion of an inference. They then use their definitions to articulate a version of IEM and the necessary conditions for falsifying it. I concede that I don’t see why their account is obligatory for investigating the connection between IEM and the sense of self-as-subject. IEM has many varieties (cf. Liang this collection, pp. 7–8 and footnote 17). In my paper (Liang this collection, pp. 2 and 6), I did not claim that the two counterexamples would undermine all versions of IEM. It was “experiential immunity” in its *de re* and *which-object* forms that came under my attack. According to experiential immunity, when I am aware of a phenomenal state through first-personal access, I cannot be wrong about whether it is me who feels it. This variety of IEM focuses on phenomenal states rather than judgments, and a key feature is that it is *relative to first-personal access*, such as introspection, somatosensation, and proprioception. This feature accommodates a widely accepted view that whether a self-ascription enjoys IEM *depends on its grounds* (Pryor 1999; Coliva 2006). The feature, however, is omitted from Haug and Jung’s account, which indicates that their version of IEM is different from my target.

Haug and Jung argue that FB’s case is not a genuine counterexample because she did not judge “I am being touched on my hand”, and hence the necessary conditions for falsifying their version of IEM are not met. However, the perplexity of this case is not why FB felt nothing when she expected that she would be touched, but why she felt the sensations when she expected that her niece would be touched. So, when FB reported feeling the sensation in test (ii), a more appropriate reconstruction of FB’s self-ascription would be: “I am being touched on my niece’s hand.” She was wrong because in fact it was her own hand being touched by the researcher, not her niece’s hand. Then, my interpretation in the paper suggested that, using Haug & Jung’s formulation, “the only reason why she was wrong was because she misidentified her own

³ I discuss the Pre-reflective Account in “Body ownership and experiential ownership in the self-touching illusion” (Liang et al. 2015).

sensations with someone else's" ([this collection](#), p. 9). This provides a falsification of experiential immunity.

Regarding the case of the body swap illusion, Haug and Jung argue that this is simply a case of mispredication. Instead of adding in more conceptual analyses to compete for the best interpretation of the study by [Petkova & Ehrsson \(2008\)](#), I will briefly describe a set of new experiments that combine the RHI and the body swap illusion. They explicitly address the Wittgenstein Question and measure the sense of experiential ownership. Before doing so, let me reply to the last issue raised by Haug and Jung.

The last issue concerns whether IEM is merely a trivial property. Here, I will limit myself to one remark. Haug and Jung consider IEM as purely a linguistic rule regarding how to use the first-person pronoun. Although many philosophers share this view, the goal of my paper was not to attack a linguistic rule. The opponents that I have in mind are those who try to use IEM to distinguish between the sense of self-as-object and the sense of self-as-subject. For these philosophers, IEM is not trivial at all. It matters to them and it matters to me if it turns out that the sense of self-as-subject really is fundamentally different from the sense of self-as-object. Because if the answer is yes, it would be very significant to consider whether the necessary and sufficient conditions for these two types of self-consciousness are distinct, and whether they are generated by different (though partially overlapping) neural mechanisms.

3 The self-touching illusion

At the end of my target paper I suggested that the next step for the investigation of the sense of self-as-subject would be to study the various conditions where one can pursue the Wittgenstein Question. I recently designed a set of experiments that allow us make exactly this step. The subject wore a head-mounted display (HMD) connected to a stereo camera positioned on the experimenter's head. Sitting face to face, they used their right hand to

hold a paintbrush, and brushed each other's left hand (figure 1).⁴ Through the HMD, the subject adopted the experimenter's 1PP as if it was his/her own 1PP. In Experiment 1, the participant watched from the adopted 3PP (180°) the front side of his/her own virtual body, including not only the torso, legs, and face, but also his/her own right hand holding a paintbrush (figure 2). In Experiment 2, the participant watched from the adopted 3PP (180°) the front side of his/her own virtual body, including the torso and legs, but not the face. The participant also saw his/her own left hand being touched by a paintbrush held by the experimenter's hand (figure 3). Compared with the asynchronous condition, the synchronous full-body condition generated a "self-touching illusion": the subject felt "I was brushing my own hand!"⁵

Two "Wittgenstein Questions" in the questionnaires were designed specifically to measure the participants' sense of experiential ownership: "It was me who felt being brushed, not someone else" (WQ1), and "The one who felt being brushed was not me" (WQ2). Notice that these two statements are directly opposed to each other. In addition, they are not about the sense of body ownership, but about *who* felt the tactile sensations caused by brushing. In Experiments 1 and 2, the participants were touched by a paintbrush, so they were indeed the subjects of those tactile sensations. This fixed the *fact* of their experiential ownership. The task was to examine whether this fact was correctly represented by their *sense* of experiential ownership. Focusing on the syn-

4 The experiments and data presented here are part of a bigger project; cf. "Body ownership and experiential ownership in the self-touching illusion" ([Liang et al. 2015](#)). Four students conducted the experiments under my supervision: Si-Yan Chang, Wen-Yeo Chen, Hsu-Chia Huang, Yen-Tung Lee.

5 The self-touching illusion was measured by two questionnaire statements: "It felt as if I was brushing my own hand" (S1), and "The one whom I brushed was me, not someone else" (S2). A Likert scale from "strongly disagree" (-3) to "strongly agree" (+3) was used for the questionnaires. In both Experiment 1 (sync. n=38, async. n=35) and Experiment 2 (sync. n=28, async. n=14), the statistics showed significant differences between the synchronous and asynchronous conditions (Exp. 1, S1: $p < 0.0010$, S2: $p < 0.0010$; Exp. 2, S1: $p < 0.0010$, S2: $p = 0.0003$; one-tailed t-test). The measurements of skin conductance responses (Exp. 1, sync. n=15, async. n=15; Exp. 2, sync. n=13; async. n=13) showed the same differences (Exp. 1, $p = 0.0080$; Exp. 2, $p = 0.0473$; one-tailed t-test). This provided objective support for the questionnaire data.

chronous conditions, the average scores on WQ1 were 1.58 and 1.04 in Experiments 1 and 2 respectively, and the average scores on WQ2 were -1.03 and -0.50 in Experiments 1 and 2 respectively.



Figure 1: Experimental set-up.

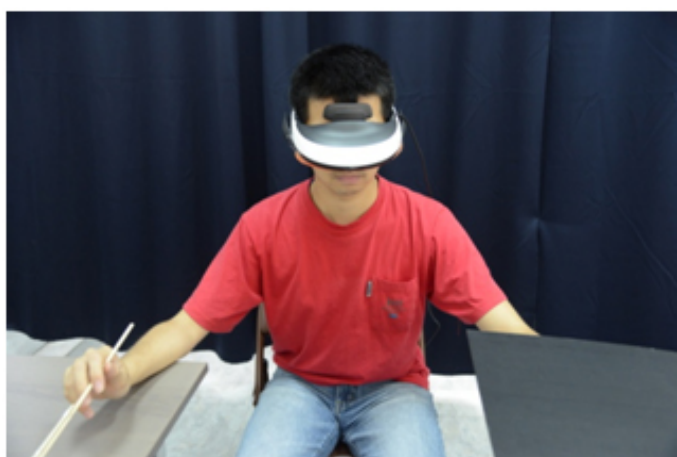


Figure 2: Subjects' view via the HMD in Experiment 1.

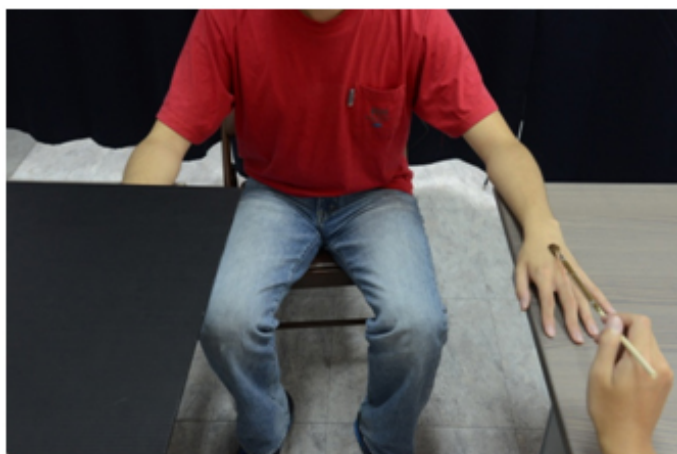


Figure 3: Subjects' view via the HMD in Experiment 2.

Suppose that the participants understood WQ1 as addressing themselves. That is, from their subjective point of view: it was *me* who felt the brushing. Then, according to IEM, no participants would commit mistakes regarding their sense of experiential ownership. One would expect that most participants would answer “strongly agree” (+3) or at least “agree” (+2) on WQ1. But that is not the case. In fact, 13.2% of participants in the synchronous conditions of Experiments 1 and 2 disagreed with WQ1 (i.e., they answered either -1, -2, or -3), and the average scores of WQ1 reported above were much lower than this interpretation requires. I discuss other possible interpretations elsewhere and argue that neither of them can support IEM.⁶ Based on the data, it is more plausible that at least some participants in these experiments were uncertain and hence prone to error about whether they were the subjects of the tactile sensations that they actually felt. That is, the fact of having tactile sensations does not guarantee that the participants will necessarily have the *sense* that “I am the one who felt them.”⁷ Overall, the data provide empirical evidence for the possibility that one’s sense of experiential ownership can misrepresent the relevant fact of experiential ownership. Hence, IEM could potentially be falsified.

⁶ Cf. “Body ownership and experiential ownership in the self-touching illusion” (Liang et al. 2015). Briefly, (i) suppose for some reason that the participants understood WQ1 to be addressing someone else. That is, in their subjective experiences, it was *not me* who felt the brushing. Then, according to IEM, one would expect that most participants would answer “strongly disagree” (-3) or at least “disagree” (-2) on WQ1. But this is not the case either. This time, the average scores of WQ1 were too high to fit this interpretation. (ii) Suppose that the participants did not all understand WQ1 in the same way: some took it as addressing themselves, but others as addressing someone else. Then, assuming IEM holds, one would expect the participants to answer either +3 (or at least +2) or -3 (or at least -2). But, again, that is not the case. Many participants answered “slightly disagree” (-1), “not sure” (0), or “slightly agree” (+1). In fact, the standard deviation in each experiment is large (Exp. 1, SD=1.5001; Exp. 2, SD=1.5512), suggesting that the participants’ responses to WQ1 varied widely.

⁷ In addition to WQ1, we also presented WQ2 (“The one who felt being brushed was not me”) in the questionnaires. The direct contrast between WQ2 and WQ1 was so obvious that, even if the participants felt uncertain about WQ1, the contrast can still be easily recognized. So, if IEM holds, one could reasonably expect that participants’ responses would manifest a *strong* “negative correlation” between WQ1 and WQ2. For example, if a subject answers +3 to WQ1, then he/she would likely answer -3 (or at least -2) to WQ2, etc. However, we only observed a weak negative correlation between these two sets of results (coefficient $R=-0.3278$).

4 Conclusion

The defenders of IEM will try to find ways to interpret these data differently. It would not surprise me if what these data mean continues to be controversial. However, I hope that experiments like these and the discussions in the target paper will at least convince many researchers that sometimes it does make sense to ask Wittgenstein Questions (like WQ1 and WQ2 above). Both the sense of self-as-subject and IEM are open to empirical as well as philosophical investigation.

References

- Bottini, G., Bisiach, E., Sterzi, R. & Vallar, G. (2002). Feeling touches in someone else's hand. *NeuroReport*, 13 (11), 249-252.
- Coliva, A. (2006). Error through misidentification: Some varieties. *Journal of Philosophy*, 103 (8), 403-425.
- Gallagher, S. (2012). First-person perspective and immunity to error through misidentification. In S. Miguens & G. Preyer (Eds.) *Consciousness and subjectivity* (pp. 187-214). Heusenstamm, GER: Ontos Verlag.
- Haug, O. & Jung, M. F. (2015). Are there counter-examples to the immunity principle? Some restrictions and clarifications: A commentary on Caleb Liang. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1-14). Frankfurt a. M., GER: MIND Group.
- Legrand, D. (2006). The bodily self: The sensori-motor roots of pre-reflective self-consciousness. *Phenomenology and the Cognitive Sciences*, 5 (1), 89-118.
[10.1007/s11097-005-9015-6](https://doi.org/10.1007/s11097-005-9015-6)
- (2007). Pre-reflective self-as-subject from experiential and empirical perspectives. *Consciousness and Cognition*, 16 (3), 583-599.
[10.1016/j.concog.2007.04.002](https://doi.org/10.1016/j.concog.2007.04.002)
- (2010). Myself with no body? Body, bodily-consciousness and self-consciousness. In D. Schmicking & S. Gallagher (Eds.) *Handbook of phenomenology and cognitive science* (pp. 181-200). Dordrecht, NL: Springer.
- (Ed.) (2010). Myself with no body? Body, bodily-consciousness and self-consciousness. In D. Schmicking & S. Gallagher (Eds.) *Handbook of phenomenology and cognitive science* (pp. 180-200). Dordrecht, NL: Springer.
- Liang, C. (2015). Self-as-subject and experiential ownership. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1-20). Frankfurt a. M., GER: MIND Group.
- Liang C., Chang, S-Y., Chen, W-Y., Huang, H-C. & Lee, Y-T. (2015). Body ownership and experiential ownership in the self-touching illusion. *Frontiers in Psychology*, 5 (1591). [10.3389/fpsyg.2014.01591](https://doi.org/10.3389/fpsyg.2014.01591)
- Petkova, V. & Ehrsson, H. (2008). If I were you: Perceptual illusion of body swapping. *PLoS ONE*, 3 (12), e3832. [10.1371/journal.pone.0003832](https://doi.org/10.1371/journal.pone.0003832)
- Pryor, J. (1999). Immunity to error through misidentification. *Philosophical Topics*, 26 (1 & 2), 271-304.
[10.5840/philtopics1999261/246](https://doi.org/10.5840/philtopics1999261/246)
- Zahavi, D. (2005). *Subjectivity and selfhood*. Cambridge, MA: MIT Press.

Mathematical Cognition

A Case of Enculturation

Richard Menary

Most thinking about cognition proceeds on the assumption that we are born with our primary cognitive faculties intact and they simply need to mature, or be fine-tuned by learning mechanisms. Alternatively, a growing number of thinkers are aligning themselves to the view that a process of enculturation transforms our basic biological faculties. What evidence is there for this process of enculturation? A long period of development, learning-driven plasticity, and a cultural environment suffused with practices, symbols, and complex social interactions all speak in its favour. In this paper I will sketch in outline the commitments of the enculturated approach and then look at the case of mathematical cognition as a central example of enculturation. I will then defend the account against several objections.

Keywords

4E cognition | Ancient number system | Arithmetical cognition | Cognitive integration | Cultural inheritance | Discrete number system | Enculturation | Evolution of cognition | Evolutionary continuity | Mathematical cognition | Niche construction | Symbol systems | Symbolic thought

Author

[Richard Menary](#)

richard.menary@mq.edu.au

Macquarie University
Sydney, NSW, Australia

Commentator

[Regina Fabry](#)

fabry@students.uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

Since cognitive science took an ecological turn it has been casting around for new frameworks in which to conduct its main business: experimental research. Those who have taken the ecological turn are convinced that classical and brain-bound frameworks don't provide the necessary conceptual and experimental tools required to make sense of cognition in the wild ([Hutchins 1995](#)). A number of alternative frameworks have been proposed, with embodied cognition the most frequently adopted. The theoretical framework one uses to understand cognition has profound empirical consequences for scientific practice. For example, it influences

what we consider to be the relevant phenomena of interest, what questions we ask about them, how we design and perform experiments, and how we interpret results ([Beer 2000](#)). The theoretical framework of classical computation, for example, approaches cognitive processing as a matter of input represented symbolically, which is then syntactically processed according to stored knowledge that the system has. It proposes a single "sandwich style" layer of cognitive processing, involving input, computation, and output ([Hurley 2010](#)).

The theoretical framework of CI (cognitive integration; [Menary 2007](#)) proposes something

altogether different: multiple cognitive layers where neural, bodily, and environmental processes all conspire to complete cognitive tasks. Although the framework is unified by a dynamical systems description of the evolution of processing in the hybrid and multi-layered system, it recognises the novel contributions of the distinct processing profiles of the brain, body, and environment. Furthermore, the CI framework explains our cognitive capabilities for abstract symbolic thought by giving an evolutionary and developmental case for the plasticity of the brain in redeploying older neural circuits to new, culturally specific functions—such as reading, writing, and mathematics (Menary 2014). I call this a process of enculturation.

This paper seeks to outline the phylogenetic and ontogenetic conditions for the process of enculturation. It will take mathematical cognition, particularly the evolutionary basis for mathematical cognition, as a core example of enculturation. In so doing, I hope to have given an account of why enculturation exists, how it happens, and in what ways it can be defended against objections. In the [first](#) section I will explore the relationship of CI to cognition embodied, embedded, enacted, extended (4E) cognition and then explain why social and cultural practices are important to the process of enculturation. In the [second](#) section I will outline the core concepts required to make sense of enculturation: continuity, transformation, novelty, and uniqueness. The [third](#) section will introduce the example of mathematical cognition, moving from the evolutionary basis for numerosity and numerical cognition to the precise operations of mathematics. The [fourth](#) section will give an account of mathematical cognition as a case of enculturation. In the [final](#) section I outline two possible objections and respond to them.

2 Where does CI sit in the 4E landscape?

Traversing the 4E landscape one rises from the lowlands of weakly embodied and embedded cognitive science to the giddy heights of strong embodiment and embedding. Embodied cognition is the thesis that at least some of our cognitive states and processes are constituted by

bodily processes that are not brain-bound. Embodied cognition is the thesis that our cognitive systems are located in and interact with the surrounding physical and social environment. Enactive and extended approaches to cognition inhabit the rarefied atmosphere of the strongly embodied and embedded peaks. However, there are important differences between enaction and extension and between those variants and CI. To determine where CI and enculturation sit in the 4E landscape, I will use a dimensional analysis I first introduced in [Menary \(2010\)](#).

Embodied mind

Embodied mind weak: the mind/brain is embodied (compatible with internalism/individualism [Smart 1959](#); [Stich 1983](#))

Embodied mind moderate: some of our mental and cognitive processes and states depend¹ upon our non-neural body ([Gallagher 2005](#); [Gallese 2008](#))

Embodied mind strong: some of our mental and cognitive processes and states are constituted by processes of the body acting in and on the environment (compatible with enactivism [Varela et al. 1991](#), and CI [Menary 2007](#))

Embedded mind

Embedded mind weak: All the perceptual inputs to and behavioural outputs from cognitive systems are found in the environment (compatible with internalism/individualism [Adams & Aizawa 2008](#); [Rupert 2009](#))

Embedded mind moderate: Mental and cognitive states and processes are scaffolded or causally depend upon the environment ([Sterelny 2003](#); [Wheeler 2005](#))

Embedded mind strong: Some mental and cognitive processes and states are integrated with environmental states and processes into a single system (compatible with extended mind [Clark 2008](#), [this collection](#); [Menary 2007](#); [Rowlands 2010](#))

¹ Here we might take dependence simply to be a causal, and not a constitutive, relation. Perhaps my gesturing in a particular way causes my recalling a word.

Weakly embodied mind is just the old thesis that the mind is identical to the brain. One can be an individualist and hold to this form of embodiment, and I won't consider the implications of the view here. The work of some² embodied cognition researchers will fall under the moderate sense of embodiment. For example, those who attempt to show that concepts or word-meanings are causally dependent upon sensorimotor areas of the brain (Glenberg 2010; Gallese 2008) commit to a moderate sense of embodiment. The strong sense of embodiment focuses on how cognition is constituted by bodily interaction with the environment, and I shall focus on the discussion here. CI and enactivism occupy this region of the environment, but with different emphases on the nature of the interaction and the evolutionary continuity of simple and complex cognitive systems. CI also occupies the strongly-embedded region, but I shall deal with the relation between CI and cognitive extension in the next sub-section.

Enactivism (excluding its radical variant)³ allows that even simple living systems are cognitive. Enactivists are committed to the continuity of life and mind and so they propose cognitive and even mental states and processes⁴ for much simpler biological systems than would CI (Varela et al. 1991).⁵ Whilst I am sympathetic with the commitment to continuity between simple cognitive systems and complex cognitive systems, it is questionable whether we should argue that simply being a living organism provides sufficient cognitive complexity for conscious experience and sense (or meaning) making.

CI does not require us to think that complex cognitive and mental phenomena, such as conscious experience, are shared by all living or-

ganisms whatever their complexity or simplicity. This is to assume that the properties of complex cognitive systems will be found even in very simple cognitive systems. According to CI, this gets things the wrong way round: there is a continuity from very simple systems that interact with their environments, by having mechanisms that track or detect salient features of their environments, to complex systems that have a wider range of cognitive capabilities (traits) including memory, inference, communication, problem solving, social cognition, and so on. By contrast a phylogeny of cognitive traits would show the distribution of those traits (across species) and help us to understand both the evolutionary pressures that produce more complex kinds of cognitive systems and the innovations that bring about new traits.⁶

CI provides a phylogenetic and ontogenetic basis for when bodily interactions are cognitive processes. Along with niche constructionists (Laland et al. 2000), CI maintains a phylogeny of hominid cognition in terms their active embodiment in a socially constructed cognitive niche. Ontogenetically, neonates acquire cognitive abilities to create, maintain, and manipulate the shared cognitive niche, including tools, practices, and representational systems. Cognitive processing often involves these online bodily manipulations of the cognitive niche, sometimes as individuals and sometimes in collaboration with others. CI has a unique position on the 4E landscape, because it is the first framework to propose that the co-ordination dynamics of integrated cognitive systems are jointly orchestrated by biological and cultural functions. What, though, are the cultural functions in question?

2.1 Cognitive practices as cultural practices

Both CI and extended mind (EM) occupy the strong embedding region, but they do so in different ways. Here I will differentiate CI as a thesis of enculturation from Clark's organism-

² One could look at a classic paper on mind/brain identity such as Smart (1959).

³ See Thompson (2007) for an account of the life-mind continuity, Stewart et al. (2010) for a volume dedicated to enactivism, and Hutto & Myin (2013) for a self-proclaimed radical variant.

⁴ See for example Barbaras (2010), which argues that to live is to have intentional consciousness of living.

⁵ Interestingly, radical enactivists appear to agree with CI on this issue; see Hutto & Myin (2013, p. 35). However, the radicals have a problem bridging the gap between basic cognitive processes and enculturated ones, since they think that meaning, or content, can only be present in a cognitive system when language and cultural scaffolding is present (Hutto & Myin 2013). That, of course, doesn't sit well with evolutionary continuity.

⁶ See for example Sterelny's cognitive phylogeny in Sterelny (2003) and Godfrey-Smith's complexity thesis in Godfrey-Smith (1996). See MacLean et al. (2012) for an overview of the problems for a comparative phylogeny.

centred approach to EM. Cognitive integration is a model of how our minds become enculturated. Enculturation rests in the acquisition of cultural practices that are cognitive in nature. The practices transform our existing biological capacities, allowing us to complete cognitive tasks, in ways that our unenculturated brains and bodies will not allow. Cultural practices are patterns of action spread out across cultural groups (Roepstorff 2010; Hutchins 2011; Menary 2007, 2010, 2012). Cognitive practices⁷ are enacted by creating and manipulating informational structures⁸ in public space. This can be by creating shared linguistic content and developing it through dialogue, inference, and narrative; or it can be by bodily creating and manipulating environmental structures, which might be tools or public and shared representations (or a combination of both). Examples of linguistically mediated action include self-correction by use of spoken (or written) instructions, co-ordinating actions among a group, or solving a problem in a group by means of linguistic interaction. Examples of creating and manipulating public and shared representations include using a graph to represent quantitative relationships; using a diagram to represent the layout of a circuit or building; using a list to remember a sequence of actions; or to solve an equation, to mathematically model a domain, to make logical or causal connections between ideas, and so on. Practices can be combined into complex sequences of actions where the physical manipulation of tools is guided by spoken instructions, which are updated across group members. A simple example of a group brainstorming with one member writing out the answers would be an example of a complex of collaborative cognitive practices.⁹

Cognitive practices are culturally endowed (bodily) manipulations of informational structures.

Practices govern how we deploy tools, writing systems, number systems, and other kinds of representational systems to complete cognitive tasks. These are not simply static vehicles that have contents; they are active components embedded in dynamical patterns of cultural practice. Practices are public, and they are also embodied and enacted.¹⁰ We embody practices: they become the ways in which we act, think, and live. They structure our lifeways (although not exclusively).

CI does not deny that much thinking takes place offline in the brain, but it does take the online and interactive mode of thought to be adaptive. Again, this line of thought has precursors,¹¹ but CI, uniquely, takes interactive thought as a basic category,¹² which is then scaffolded by culturally evolved practices. Practices stabilise and govern interactive thought across a population of similar phenotypes. The stable patterns of action can then be inherited by the next generation, because the practices have become settled and are part of the developmental niche in which the minds of the next generation grow. Our brains co-adapted to the stable spread of practice and its role in ontogeny—resulting in the slow evolution of the cultural brain.

The focus upon practice and culture marks cognitive integration out from variants of extended cognition, such as Clark's organism-centred approach to extension (2008). Clark's organism centred approach takes the assembly of extended cognitive systems to be controlled by the discrete organism, and brain, at the centre of it. He thereby reduces the role of cultural practices in large or small groups of organ-

⁷ I don't mean to suggest that there can't be other effects of cognitive practices, but since practices are just the cultural formalisation of patterns of action across a population, or group, cognitive practices are tied directly to these patterns of action. I can't provide a detailed origin account for cognitive practices here, but see Menary (2007, Ch. 5) for an early attempt to do so. However, the account of mathematical cognition I give in the next two sections provides an example of how such an account would be likely to look.

⁸ The primary cases I am thinking of are public systems of representation, including spoken language. However, I don't want to rule out cases involving tools, bodily gestures, artistic or bodily adornments, and the intelligent use of space and objects.

⁹ For two very good overviews of collective or group cognition see Theiner (2013) and Huebner (2013).

¹⁰ Jennifer Windt helpfully pointed out that practices can be thought of as public, because they are embodied and enacted. I think that this is just right: practices are patterns of action spread across a population. However, I am inclined to think that practices are not simply reducible to the bodily actions of individuals. Whilst doing long multiplication requires a bodily action of me, what I am doing cannot be described exclusively in terms of those bodily actions. The practice is a population, or group level phenomenon, not an individual one.

¹¹ The classical pragmatists, particularly Peirce and Dewey, held that thought was interactive. See Menary (2011) for a description of pragmatist approaches to thought, experience and the self.

¹² See Menary (2007, Ch. 5), where I make a detailed evolutionary case.

isms in the explanation of cognitive assembly. “Brains are special, and to assert this need mark no slippery-slope concession to good old-fashioned internalism as an account of mind. It is fully consistent with thinking (as I do) that Hutchins is absolutely right to stress the major role of transmitted cultural practices in setting the scene for various neurally-based processes of cognitive assembly” (Clark 2011, p. 458). On Clark’s view, cultural practices only set the scene for the real work of integration to be done by the brain. Whilst it is arguable whether Clark’s position is a return to “good old fashioned internalism,” he certainly does not give cultural practices a central role in assembling and orchestrating cognitive systems.¹³ Hutchins, by contrast, is committed to a full-blooded enculturated approach:

[t]he ecological assemblies of human cognition make pervasive use of cultural products. They are always initially, and often subsequently, assembled on the spot in ongoing cultural practices. (2011, p. 445)

CI is the only variant of strong embedding (including EM) to explain the role of cultural practices in assembling integrated cognitive systems. Cognitive practices are inherited as part of the developmental niche and have profound transformative effects on our cognitive abilities. This leads us to the main concepts required to understand these transformations as a process of enculturation.

3 Enculturation: The main concepts

In this section I define and explain the main concepts required to understand enculturation, other than the already explored concepts of integration and practice. I will develop the concepts of evolutionary continuity, behavioural and neural plasticity, transformation and innov-

¹³ If this is an accurate portrayal of Clark’s position (and I have tried to carefully use his own words) then, despite his protestations to the contrary, it appears to be a return to internalism, at least for the most central and important cognitive processes. If the brain carries out all the important cognitive operations, then Clark’s position would be a moderate embedded cognition for core cognitive abilities and an extended approach only to some of the more peripheral cases.

ation, or novelty and uniqueness. In particular I will emphasise the phylogenetic and ontogenetic bases for modern human cognitive capacities.

3.1 Evolutionary continuity

The concept of evolutionary continuity results from the fact that evolution occurs gradually with complex structures evolving over many generations. Over long periods of time these gradual changes accumulate, resulting in large differences. Consequently, changes to a phenotype occur in slow cumulative steps over long periods of time and do not appear in a single mutational step. Evolutionary continuity demands that modern human minds evolved from earlier archaic variants. Doubtless modern minds differ from archaic minds in important respects, but these differences must have evolved over long periods of time, through slow cumulative mutational changes to the genotype. Even so, we should expect some of our archaic traits to remain, and for more modern variants to be built on top of them. One obvious example of this is the evolution of the human brain.

The evolution of the human brain can, to some extent, be seen in the gradual increase of cranial capacity, but some of the most important changes have been in the reorganisation of cortical circuitry and interconnectivity (Hoffman 2014). Although the evolution of the human brain can be understood in terms of increasing encephalization and increased connectivity between brain regions, the human brain has essentially the same set of structures as any other primate brain.¹⁴ Modern brains evolved from archaic brains and share the same evolutionary constraints as other primates: “the similarity in brain design among primates, including humans, indicates that brain systems among related species are internally constrained and that the primate brain could only evolve within the context of a limited number of potential forms” (Hoffman 2014, p. 5). Modern minds are still partly archaic.

¹⁴ “Although species vary in the number of cortical areas they possess, and in the patterns of connections within and between areas, the structural organization of the primate neocortex is remarkably similar” (Hoffman 2014, p. 4).

It is important to think of evolutionary continuity as running from archaic to modern. We should try to avoid anthropomorphic tendencies to project modern cognitive capacities backwards into the hominin lineage or across to primate species. For example, humans are excellent social cognisers, but it does not follow from this that we should expect other primates to have a theory of mind.¹⁵ The evolutionary pressures under which humans evolved and the capacities for complex social cognition might have been very different from those under which other primates evolved. Consequently, we should be searching for archaic precursors to modern cognitive capacities. For example, we might expect that given the increasing social pressures in hominid social groups there would be precursors to modern social cognition and that these precursors would have been adaptive solutions (Shultz et al. 2012). Modern human social cognition would then be an evolutionary consequence of increasing variation in the complexity of social organisation and interaction (Sterelny 2003).

I am committed to another sense of continuity: that between biology and culture. Culture is not, as a category, distinct from the biological. Although culture is sometimes thought of as floating free of our biological nature and sometimes as being highly constrained by it, I shall assume that genes and culture co-evolve¹⁶ mutually, influencing and constraining one another. Therefore I shall accept no culture–biology dualism in this paper. Indeed I shall adopt a cultural inheritance model of cognitive evolution (of the niche construction kind). However, I shall always do so with archaic origins in mind. Archaic origins matter to cognitive evolution and they matter to the way our brains develop during the lifespan.¹⁷

¹⁵ Indeed, it is questionable whether humans deploy a theory of mind, or at least, perhaps they only do so on rare occasions (Hutto 2008; Andrews 2012). Andrews has also argued that we may share a number of “mind reading” strategies with other primates that don’t involve theory of mind (2012).

¹⁶ See below for a niche construction account of gene–culture co-evolution. I favour such an account because it helps us to understand how a developmental niche could have cumulative downstream evolutionary effects on phenotypes (Sterelny 2003).

¹⁷ They matter because they are part of the developmental biases that produce a robust phenotype.

In the “modern synthesis” there is only one line of inheritance, and that is genetic inheritance. More recently, biologists (Odling-Smee et al. 2003) have proposed that there are other lines of inheritance: ecological inheritance and cultural inheritance (Boyd & Richerson 2005). Many organisms construct the niche in which they live, mate, hunt, and die. Niche constructors modify the ancestral environment, and these modifications are bequeathed to the next generation. Modifications encompass physical alterations, such as living in mounds or constructing hives, as well as cultural artefacts, practices, and institutions. Over long periods these alterations to the niche can have profound effects on the phenotype. For example, the ubiquitous niche constructions of termites, burrows and mounds, have profoundly altered their morphology and behaviour (Turner 2000).

Humans are also ubiquitous niche-constructors. They physically alter their environment and they also epistemically, socially, and culturally engineer the environment (Sterelny 2003, 2010; Menary 2007). Humans are born into a highly structured cognitive niche that contains not only physical artefact, but also representational systems that embody knowledge (writing systems, number systems, etc.); skills and methods for training and teaching new skills (Menary & Kirchhoff 2014); and practices for manipulating tools and representations. Inherited cultural capital is a real and stable feature of the socio-cultural environment, including a great variety of knowledge systems, skills, and practices across a variety of domains of human action. As such, human cultural niches provide neonates with rich developmental niches. It is in these developmental niches that humans acquire cognitive practices.

Cognitive practices are products of cultural evolution, evolving over faster timescales than biological evolution. Writing systems, for example, are only thousands of years old; consequently, it is highly unlikely that there is a “reading gene” or even an innate specialised “reading module.” This is important: cognitive capacities for reading and writing, mathematics, and other culturally recent forms of cognition could not be biological adaptations (that

evolved over long periods of time). The timescales for their evolution are too short. It follows that the capacity for culturally recent forms of cognition must be acquired through learning and training.

Although there are no innate specialized modules for these recent forms of cognition, cortical circuits with which we are endowed through evolution are transformed to perform new culturally recent cognitive functions, even though they evolved to perform different functions. Recent cognitive innovations aside, there are good reasons to expect that evolution has driven us to think by interacting with the environment and that this is adaptive (Sterelny 2003 2012; Menary 2007; Wheeler & Clark 2008). However, it is the scaffolding of cultural practices that orchestrates the interactions—as in the case of written language and mathematics.

Structured socio-cultural niches have had profound evolutionary consequences in the hominin lineage. Structured niches have co-evolved with human phenotypic and developmental plasticity. We have evolved to be a behaviourally plastic species (Sterelny 2012) as well as a cultural species. In this co-evolution we have developed all manner of skills, practices, and activities. Why, though, are we so peculiarly behaviourally plastic? One good answer to this question is that human behavioural and developmental plasticity is an adaptive response to the variability and contingency of the local environment (Finlayson 2009; Sterelny 2003, 2012; Davies 2012). This is an alternative to the view that we are adapted to a pleistocene hunting and gathering environment—a view relied upon by many evolutionary psychologists (Barkow et al. 1992).

Critical to a co-evolutionary account of cultural practices is the evolution of human plasticity. Given that there is such a variety of cultural activity, we need an account of human evolution that will allow for variability in human behaviour. Second, we need a model that explains how innovations in our cultural niche are inherited and propagated, leading to changes in behaviour over time. The niche construction model explains how both of these causal factors could come into play. In the sub-

sections below, I outline the importance of behavioural and neural plasticity, the concept of transformation, and those of novelty and uniqueness.

3.2 Behavioural and neural plasticity

In evolutionary terms, humans are capable of developing a wide range of skills that allow them to cope with a wide variety of environments (and their contingencies). For example, even where skills are (broadly) of the same type, such as hunting, they will vary in how they cope with the differences in local environments—think of the differences in environments between Aboriginal hunters in the Pilbara desert, hunter-gatherers in the Central American rainforests, and Inuit seal-hunters (Sterelny 2003, p. 167).

Development is extended in modern humans relative to other species. Humans take a long time to learn how to walk and talk, and much, much longer to develop fine-grained manual and cognitive skills such as reading and writing. Other primates have much faster developmental timescales. While this might make humans more dependent on their caregivers for longer, it also allows them to refine skills and acquire a greater array of them before entering adulthood.

Through cultural inheritance, knowledge, skills, and artefacts are passed on to the next generation, but learning environments and learning techniques are also passed on so that the next generation can acquire and be transformed by the inherited cultural capital. This last point is important for our purposes, because developmentally plastic humans need scaffolded learning environments in which to develop.¹⁸

How, though, are we capable of acquiring these new cultural capacities in development? Through neural plasticity. Rather than the process of synaptogenesis or lesion-induced plasticity,¹⁹ the kind of plasticity I will discuss here is

¹⁸ If the cognitive abilities for manipulating artefacts and representations are not innate, then a scaffolded learning environment helps to explain how we acquire them.

¹⁹ Many neurological studies of plasticity focus on synaptogenesis, the florid growth of grey matter and then the consequent pruning, or the

what I call learning driven plasticity (see [Menary 2014](#)). Learning driven plasticity (LDP) can result in both structural and functional changes in the brain. Structurally, LDP can result in new connections between existing cortical circuits. Functionally, LDP can result in new representational capacities (the ability to represent public symbolic representations such as alphabets and numerals) and new cognitive abilities, such as mathematics,²⁰ reading, and writing ([Dehaene 2009](#); [Ansari 2012](#)). It should come as no surprise that learning drives structural and functional changes in the brain, given the extended developmental period in humans and the late development of the cortex ([Thatcher 1991](#)). The brain changes, not just because of maturation, but also because of learning:

[w]hen children learn to read, they return from school ‘literally changed’. Their brains will never be the same again. ([Dehaene 2009](#), p. 210)

Famously, Dehaene argues that a region of the occipito-temporal junction (which he calls the VWFA, visual word form area) that is part of a wider network for recognising faces, objects, and even abstract shapes (such as chequer patterns), alters its function to recognise written symbols in alphabets and even logographic scripts such as kanji ([Dehaene 2009](#)). This is due to the plasticity of that area of the brain, where the functional shift is due to scaffolded learning.²¹ “Scanning of ‘ex-illiterate’ adults who learned to read during adulthood has demonstrated that the VWFA is highly plastic, even in adults, and quickly enhances its response to letter strings as soon as the rudiments of reading are in place” ([Dehaene & Cohen 2011](#), p. 259). Even those who are not convinced that a specialised region for “word recognition” is acquired once we learn to read admit that the oc-

cipito-temporal junction is part of a reading and writing circuit (e.g., [Price & Devlin 2011](#)).

We have evolved to be phenotypically and developmentally plastic. This is in no small part due to the plasticity of our brains. Our developmentally plastic brains exhibit learning-driven plasticity. When the brain is coupled to a highly scaffolded learning environment it is profoundly transformed, structurally and functionally, and consequently we are cognitively transformed in the profoundest way.

3.3 Transformation

The transformation thesis can be given a simple formulation: cognitive transformations occur when the development of the cognitive capacities of an individual are sculpted by the cultural and social niche of that individual. Cognitive transformations result from our evolved plasticity and scaffolded learning in the developmental niche. In the previous sub-sections an account was given of the effects of cultural inheritance and niche construction on hominid evolution. The result is phenotypic plasticity, and in the cognitive case the co-evolution of neural plasticity and scaffolded learning. However, the point of the transformation thesis is to drill down into the process of acquiring knowledge, skills, and cognitive abilities via learning-driven plasticity and scaffolded learning. It does this by showing how transformations are a result of the role of cognitive practices in development. Practices structure the niche; they transform plastic brains via learning driven plasticity and result in new cognitive abilities.

During the learning and training of a skill, such as flaking an arrowhead, or a shot in tennis or cricket, we are guided by the norms for the correct actions that make up the skilled practice. A parallel case can be made for cognitive abilities such as mathematics. The neophyte mathematician gains mastery over the cognitive norms²² by which numerals, operators, and other symbols are created and manipulated. Vygotsky expresses this in the claim that children, “master the rules in accordance with which ex-

synaptic death of many of those neurons in the so-called critical period of childhood. There are a large number of studies of neural damage, often by stroke or injury, where cortical circuitry becomes damaged and its function impaired, but where other areas of the cortex can take on the impaired function. (See [Huttenlocher 2002](#) for an overview.)

²⁰ I will be defending an account of mathematical cognition in section 4.

²¹ See [Menary \(2014\)](#) for a discussion of plasticity and the VWFA.

²² For an account of cognitive norms see [Menary \(2007\)](#), Chapter 6.

ternal signs must be used” (Vygotsky 1981, pp. 184–185). Initially the child masters the creation and deployment of spoken linguistic signs (and later written signs) through the scaffolding of parents and caregivers. However, this process is not simply a matter of gaining new representations; it is also one of gaining new abilities.

Neophytes go through a process of dual-component transformation: they learn how to understand and deploy public symbolic representations and they learn how to create and manipulate inscriptions of those symbols in public space (Menary 2010). In so doing, they learn mathematical and linguistic concepts and they learn how to manipulate inscriptions to complete cognitive tasks. When learning the manipulative techniques, the first transformation is one of the sensory-motor abilities for creating and manipulating inscriptions: we learn algorithms like the partial products algorithm²³ and this is an example of the application of a cognitive practice. This is something we learn to do on the page and in the context of a learning environment, in public space, before we do it in our heads. Our capacities to think have been transformed, but in this instance they are capacities to manipulate inscriptions in public space. This is a way of showing that the transformation of our cognitive capacities has recognisably public features. This ought not to be a surprise, given that the cognitive niche is socially and culturally constructed and is structured by socio-cultural practices. Symbol systems, such as those for written language and mathematics, are not impermanent scaffolds that we shrug off in adulthood, but are permanent scaffolds that indelibly alter the architecture of cognition.²⁴

The transformatory position is quite different from that held by Clark or Sterelny. In particular it holds that our basic cognitive capabilities are transformed in development and that the dual component transformation results in a distinct functional redeployment of neural circuitry and new abilities to bodily manipulate structures in public space. Cognitive tasks can be completed by manipulating written symbols in public space or by off-line strategies for completing algorithms, or a combination of both.

²³ I’ll look at this example in detail in section 5.

²⁴ I take this issue up again in section 4.1.

This conclusion sits happily with the idea that thought is interactive and governed by practices.

The main difference between the position outlined here and Clark’s (e.g., 2008), is that Clark does not explain cognitive extension in terms of the transformation of basic cognitive resources during development in a socio-cultural niche (although he does acknowledge the importance of symbolically structured niches). Rather, he thinks that basic biological resources are not really transformed but simply dovetail to external symbols (Clark 2008, 2011). Sterelny (2010) concentrates on cognitive scaffolding, but does not think that the manipulation of symbols in public space is constitutive of cognitive processing. The enculturated approach of CI answers questions that are problematic for both Clark and Sterelny:

1. How do we learn to complete cognitive tasks that require the manipulation of symbols in public space?
2. Assuming that cognitive processing criss-crosses between neural space and public space, how does it do this?

The first question is hard for Clark since he does not think that our basic cognitive resources get transformed, at least in the way that I have presented here. The second question is hard for Sterelny because he limits himself to a scaffolded view of cognition rather than an extended view. Consequently, manipulations of symbols in public space are not cognitive processes for Sterelny.²⁵

CI as a process of enculturation requires a robust transformation thesis. A robust transformation thesis is warranted by phenotypic and neural plasticity, in particular by learning driven plasticity. Novel and unique public systems of representation drive the transformation of our existing cognitive abilities.

3.4 Novelty and uniqueness

Sometimes symbols and tools provide us with novel functions: they radically extend our cap-

²⁵ Or they might be assuming that Sterelny does not care either way; in private communication Sterelny indicated that he does not think that boundary disputes are of much interest.

abilities in some sphere. Take the humble hand axe. Very crude hand tools have been discovered dating as far back as 2.6 mya (million years ago; [Toth & Schick 2006](#)), since then there has been evidence of a hominid capacity for cumulative cultural inheritance “which was ultimately to transform *Homo sapiens* into the richly cultural species we are today” ([Whiten et al. 2011](#)). However, the capacity for developing novel functions and transmitting them to the next generation with high fidelity appears to be a more recent innovation, as evidenced by the long periods of relative stability in technological development in the early hominids and archaic humans. It also appears to be an innovation unique to the homonin lineage ([Whiten et al. 2011](#)). The Oldowan period begins in the lower paleolithic with *Homo Habilis* around 2.6 mya, being taken up by *Homo Erectus* and *Ergaster* and ending at about 1.8 mya ([Lycett & Gowlett 2008](#)). The tool types and process of manufacture remain consistent during this period, with some refinement and novelty ([Lycett & Gowlett 2008](#)), where the main tool types were choppers and scrapers or mode 1 tools ([Semaw et al. 2003](#)).

Homo Habilis is unique in that it is the first hominid to make tools that were made to endure and be re-usable (it is likely that earlier anthropocines used naturally-occurring objects as tools that were disposable; [Jefferies 2010](#)).

Oldowan toolmaking involves the production of sharp-edged flakes by striking one stone (the core) with another (the hammerstone). Effective flake detachment minimally requires visuomotor coordination and evaluation of core morphology (e.g., angles, surfaces) so that forceful blows may reliably be directed to appropriate targets ([Stout et al. 2008](#), p. 1940).

There is a clear transition to Achulean technology at around 1.7 mya with the appearance of *Erectus*/*Ergaster*. The main innovation for Achulean technology was the bifacial handaxe—a handheld cutting tool with two cutting sides. The real explosion in novelty occurs in the upper paleolithic period, from 50,000 years ago (ya) to 10,000 ya (or to just before the advent

of agriculture and the neolithic period), with genuine novelty in tool production and use and cultural diversification. In this period we begin to see evidence of art, including paintings and sculpture, fishing, jewellery, burial, evidence of musical activity, and all the hallmarks of behaviourally modern humans. It is in this period that the combination of inherited cultural capital, with phenotypic and learning-driven plasticity, complex social relations and language results in an explosion of cultural and behavioural diversity.

It is also in this period that we begin to find evidence of proto-numerical and writing systems as novel representational innovations. Simple tally notch systems on bone fragments have been dated to between 35,000 and 20,000 ya, and may have been used for a variety of purposes, the most obvious being to keep track of economic exchanges. However, it is far easier and more economical to keep track of larger amounts using a single symbol, rather than a one-to-one correspondence of marks with things.

The complex social and economic pressures that required tracking exchanges involving increasingly large numbers would be the kind of socio-economic pressures that produced symbolisation of quantity. Social and cultural pressures can drive evolutionary novelty, in this case symbolisation and uniqueness—symbolic representations are unique in both type and property, no other animal produces written symbols to represent concepts. Symbols have unique properties that allow for operations—addition, subtraction, multiplication, division, and so on that are much harder (if not unlikely) without them.

Early symbolic number systems date from between 3000–4000 BCE, but genuinely abstract symbol systems are even more recent—about 1000–2000 BCE. The invention of symbol systems is too recent to be a genetic endowment, but is inherited as cultural capital and acquired through high-fidelity social learning (which is in turn dependent upon neural plasticity).

The phylogeny of hominid tool-use is one of hard-won innovation and retention. Modern humans have developed high-fidelity modes of transmitting cultural capital vertically and horizontally. The socio-cultural pressures that led to

humans innovating symbolic representational systems are unique and very recent. Fortunately, modern human minds are flexible enough to both innovate and reliably acquire those innovations in ontogeny.²⁶ This flexibility makes modern human minds unique, and in the case of mathematical cognition unique amongst all our primate relatives.

The next section outlines mathematical cognition as a case of enculturation, and there I will explore the example of mathematical cognition by deploying the concepts refined in the first two sections.

4 Numerical cognition

In this section I outline the phylogenetic basis of mathematical cognition. That basis is in our shared sense of quantity and our ability to estimate the size of small sets by making approximate judgements of the size of the set. This ancient endowment is the basis for our mathematical competence, but it is not all there is to mathematical cognition. This is because precise mathematics depends upon a very recent and acquired public system of exact and discrete mathematical thinking. The ancient system is analogue and approximate, but mathematics requires digital and discrete representations and exact operations. These are, of course, recent additions to inherited cognitive capital. I shall show why mathematical cognition requires our ancient capacity for numerosity and how it is constituted by cognitive practices—which transform our cognitive abilities, resulting in novel and unique modern human cognitive capacities. However, this transformation results in two partially overlapping systems—the approximate number system and the discrete number system—with the latter having unique properties acquired from cultural innovation. One of the puzzles is how it is possible to move from an inherited approximate system to an acquired exact system. The process of enculturation provides the mechanisms by which such a move takes place, from the ancient capacity for numerosity to development in a socio-cultural

niche, and the orchestrating role of practices in the assembly of the cognitive systems responsible for mathematical cognition.

4.1 Numerosity in animals and humans

There is strong evidence to suggest that we have a basic analogical and non-linguistic capacity to recognise quantity and number. I think that there is overwhelming evidence for an ancient evolutionary capacity to discriminate cardinality, and to determine in an approximate way the quantity of membership of sets. It is obvious how this capacity, for only very small sets, would be beneficial for activities such as foraging, hunting, and so on.

Recent studies have revealed that the neural populations that code for number are distributed in the intraparietal sulcus (Dehaene & Cohen 2007). A growing number of studies show that both animals and humans possess a rudimentary numerical competence, which is an evolutionary endowment. For example, red-backed salamanders have been shown to choose the larger of two groups of live prey (Uller et al. 2003). Single neuron activation studies in rhesus monkeys (Nieder et al. 2006) discovered that individual neurons respond to changes in number when presented visually (and non-symbolically). These neurons are also located in the intraparietal sulci, indicating a probable cross-species homology. The neurons peak at the presentation of a specific quantity of dots, but then decrease as the numbers presented differ from the original. So a neuron that peaks at the presentation of two dots responds less to three or four dots. The further the numerical distance of the array of dots is from the magnitude to which the neuron is tuned, the lower the firing rate of the neuron. Therefore, the ancient capacity for numerosity is an approximate function, not a discrete one (DeCruz 2008).

This is not yet counting; counting is exact enumeration. Subitizing is the ability to immediately recognise the size, or number, of a small set—usually <4 . Most animals subitize, rather than count. Infant humans also appear to be able to subitize (Rouselle & Noël 2008). This ancient or approximate number system (ANS)

²⁶ This section has put together a case for the flexibility of modern minds and the ability to acquire cultural innovations quickly and easily in ontogeny.

is a non-linguistic continuous representation²⁷ of quantities above 4; Dehaene calls it the number sense (1997). Take the following example. Whilst it is easy enough to determine which of the following two boxes contains the larger number of dots without having to count them:

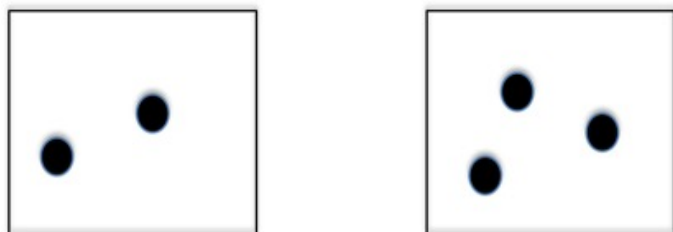


Figure 1: Subitizing or counting?

It is less easy to do so for the following (you will probably need to resort to counting):

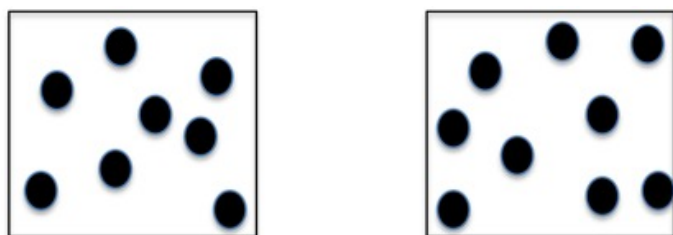


Figure 2: Subitizing or counting?

It is also possible to make estimations or approximate judgements of scale for numbers. Most people can quickly identify that 7 is larger than 3. Even for more complicated exact operations we can do this:

$$34 + 47 = 268 \text{ (is this right?)}$$

We readily reject this result, because the proposed quantity is too distant from the operands of the addition (Dehaene 2001, p. 28).

$$34 \times 47 = 1598 \text{ (is this right?)}$$

Approximation involving proximity and distance will not help here (unless you are very practised at mental multiplication), but you

²⁷ The appearance of the word representation here need not raise concerns; these are not representations with propositional contents and truth conditions. They are not symbolic and are not molecular constituents that can be combined to make more complex representations.

might resort to a multiplication algorithm (which might be routinized). It is clear that we have an ancient sense of quantity and are good at making judgements about more than and less than, but when it comes to precise and discrete quantities (particularly larger numbers) we need new capacities to be able to make judgements about operations on discrete numbers.

4.2 Two overlapping systems

The approximate numerical system is an analogue and approximate system for discriminating non-symbolic numerosities greater than 4, but the “representations” are approximate and noisy. The second system is acquired and concerns discrete symbolic and linguistic representation of individual numbers from our numeral system, including individual words for numbers. This system works with discrete, exact, symbolic representations of quantity and allows for the exact operations of arithmetic and mathematics. I will call this the discrete numerical system (DNS). There is disagreement about how much the two systems overlap. However, what is clear is that the internalisation of the public numeral system allows us to perform the kind of digital mathematical operations that are required for most arithmetic and mathematical operations (Nieder & Dehaene 2009, p. 197).

Dehaene and colleagues produced a series of experiments that demonstrate the separate functioning of the two systems. Russian–English bilinguals were taught a set of exact and approximate sums of two digit numbers in one of their languages (Dehaene et al. 1999, p. 970). Their tasks were split into giving exact answers to additions and giving an approximate answer to the addition task. The interesting result was that:

[w]hen tested on trained exact addition problems, subjects performed faster in the teaching language than in the untrained language, whether they were trained in Russian or English. (Dehaene et al. 1999, p. 971)

This provided evidence that knowledge of arithmetic was being stored in a linguistic format,

and that there was a switching cost between the trained and untrained languages. By contrast, there was equivalent performance in the approximation task, and no switching cost between the trained and untrained languages. Dehaene et al. conclude that this provides “evidence that the knowledge acquired by exposure to approximate problems was stored in a language-independent form” (1999, p. 971).

This leads us to the conclusion that there are two overlapping, but not identical, systems for mathematical cognition. The first is the ancient and approximate system, the second is a relatively new and acquired system for discrete and digital representations and operations. As Dehaene & Cohen put it:

The model that emerges suggests that we all possess an intuition about numbers and a sense of quantities and of their additive nature. Upon this central kernel of understanding are grafted the arbitrary cultural symbols of words and numbers [...]. The arithmetic intuition that we inherit through evolution is continuous and approximate. The learning of words and numbers makes it digital and precise. Symbols give us access to sequential algorithms for exact calculations. (2007, p. 41)

The two systems are overlapping but not identical because they have quite different properties. First, the ancient system is part of our phylogeny, whereas the discrete system is an acquired set of capacities in ontogeny. Second, the ancient system is analogue and approximate, whereas the discrete system is digital and exact. Third, the discrete system operates on symbols that don't map directly on to the ancient system.

When we consider very large numbers, such as 10,000,000, there is no obvious analogue in the ANS. Consequently, large or exotic numbers and operations on them do not map onto existing cortical circuitry for numerosity. Lyons et al. (2012) call this phenomenon “symbolic estrangement”. Symbols become estranged through a process of symbol-to-symbol mappings, rather than symbol-to-approximate-quantity mappings (Lyons et al. 2012, p. 635).

However, there appears to be a point of contention here: Dehaene expects there to be a more or less direct mapping of symbols to quantities (e.g., the mental number line). If symbolic estrangement does happen, then this would appear to be mistaken. Lyons, Ansari and Beilock propose a developmental resolution of this apparent disagreement. Children may start out in the acquisition of discrete number systems by a mapping to an existing approximate neural coding of quantity, but as the system matures and symbols become abstracted from the ancient system, the mature system splits into two (related but not entirely overlapping) systems: neural circuitry in the DNS tunes for discrete symbols,²⁸ whereas circuitry in the ANS tunes for approximate quantities, such that discrete symbols do not map directly onto approximate quantities. E.g., 10,000,000. The DNS has properties that are unique.

In the next section I return to the question of the role of practices in assembling the DNS.

5 Mathematical practices

The DNS is dependent upon mathematical practices, systems of number and algorithms for performing mathematical operations, complex mathematical concepts such as sets, functions, and so on. None of these practices, representations, or concepts are innate, and no one seriously thinks that they are. They are culturally inherited and acquired in the right learning niche with experts willing to teach. These new abilities are continuous with our cognitive phylogeny. How, though, can we put the whole package together? This section does that job.

5.1 Cognitive practices and the development of mathematical competence

Mathematics and writing systems are examples of culturally evolved symbol systems that are deployed to complete complex cognitive tasks. These systems are structured by rules and

²⁸ There is evidence of narrower tuning curves for Arabic numerals in the left intraparietal sulcus (Ansari 2008).

norms, but they are deployed as practices: patterns of action spread out across a population. In this case cognitive agents must gain mastery over the symbols, including numerals and operators, as well as the rules for their combination. However, they must also learn how to write and manipulate the symbols according to those rules in order to produce the right products—and this is proceduralised.

There may be more than one way of achieving a solution to the task. One can multiply by the partial products algorithm, or one can use the lattice/grid method or a number of others that have been developed by different cultures using different numerical systems. However, they all involve the same set of features: symbols, rules, operators, spatial configuration, and products, and they jointly constitute a practice for manipulating the symbols to complete mathematical problems. The practices are novel and unique to humans.

The methods apply equally to their off-line equivalents, so in the page-based version of the partial products algorithm we perform the multiplications from right to left and write down their products in rows, carrying numbers where necessary. In the off-line version we can perform the same operations on imagined numerals, multiplying numbers along the line and carrying any numbers as required. It is cognitively taxing to hold the products of the multiplications constant in working memory, though some people can train themselves to become quite good at it. Most people learn off-line multiplication by performing shortcuts; if I want to work out what 25×7 is, I just add 25 together 7 times.

On-line methods can change even within the same arithmetical systems, so the partial products algorithm works like this:

$$\begin{array}{r}
 23 \\
 \times 11 \\
 \hline
 23 \quad (1 \times 3 \text{ and } 1 \times 2) \\
 + 230 \quad (\text{carry } 0, 1 \times 2 \text{ and } 1 \times 3) \\
 \hline
 253 \quad (\text{add products together})
 \end{array}$$

However there is an equivalent algorithm that works like this:

$$\begin{array}{r}
 23 \\
 \times 11 \\
 \hline
 200 \quad (10 \times 20) \\
 30 \quad (10 \times 3) \\
 + 23 \quad (1 \times 23) \\
 \hline
 253 \quad (\text{add products together})
 \end{array}$$

The algorithms may differ, but they still involve the practice of spatially arranging the numerals, and performing operations on them and deriving a product, by performing the staged manipulations on the page. It appears then to matter how we manipulate symbols in public space, but is there any empirical evidence for this conclusion?

CI predicts that it matters how symbols are spatially arranged when they are being manipulated. Landy & Goldstone (2007) found that college-level algebraists could be induced to make errors by altering the layout of numbers that they were to manipulate. They did this by altering the spacing of the equations:

$$F+z * t+b = z+f * b+t$$

Although minor, the extra spacing was enough to induce errors. It matters how the symbols are spatially laid out, for this layout is the basis of how we manipulate those symbols. In this case the artificial visual groups created by the irregular spacing affected the judgement of the validity of the equation. If the visual groupings were inconsistent with valid operator precedence then they negatively affected the judgement.²⁹

Landy & Goldstone's work provides evidence that expert algebraists are practised at symbolic reasoning achieved via the perception and manipulation of physical notations (2007; Landy et al. 2014). Rather than an internal system of abstract symbols and rules for their combination (i.e., a language of thought), the system is composed of perceptual-motor systems and the manipulations of numerals. They are careful to say that the manipulations must conform to the abstract norms of algebra. Dutilh Novaes (2013) takes this to be evidence that mathematical competence is constituted by the

²⁹ In algebra multiplications are made before additions. E.g., $5+2*6 = 17$ (not 42).

capacity to manipulate inscriptions of mathematical equations. This fits very well with the CI approach.

Despite some interesting lacunae (savants and blind mathematicians), most mathematicians learn to manipulate numerals and other mathematical symbols on the page, and they continue to do so throughout their mature cognitive lives. Landy and Goldstone's evidence supports the thesis that mathematical competence is constituted, in part, by our capacity to manipulate symbols in public space; that competence is, properly, a matter of interaction.

5.2 Continuity and transformation

We have seen that there is an ancient evolutionary endowment for numerosity—an analogue and approximate system. This system is found in other primates and other species. It provides both the phylogenetic basis of mathematical cognition and the initial constraints for the development of the DNS. The DNS did not spring *sui generis* into the world. It did so because of a heady mixture of socio-cultural pressures, phenotypic and neural plasticity, social learning strategies, and cultural inheritance. These are the conditions for the scaffolding of the ANS, transforming our basic biological capacities into the DNS.

New cultural functions, discrete mathematical functions, and the practices for manipulating inscriptions transform existing circuitry in the brain. Once we learn how to recognise, understand, and manipulate mathematical symbols our brains undergo a profound transformation. There is a reproducible circuit for mathematical cognition involving a bi-lateral parietal based approximate estimation; a left lateralised verbal framework for arithmetic concepts (e.g., number words); and a occipito-temporal based symbol recognition system (e.g., Arabic numerals). The system also incorporates visual-motor systems for writing (manipulating, or pushing) symbols in public space.

A further important aspect of transformation is symbolic estrangement. As the DNS matures it becomes more abstract and less directly mapped onto the approximate functions of the

ANS. Interestingly, at the same time expert mathematicians become reliant upon visual-motor capacities for manipulating inscriptions. Transformation depends upon the novelty and uniqueness of mathematical symbols and practices.

5.3 Novelty and uniqueness

Symbolic number systems and sequential algorithms allow for mathematical and cognitive novelty. Once we have a public system, all manner of exotic numbers and operations can be discovered:³⁰ negative numbers, square roots, zero, sets, and so on. Its importance lies in the ability to perform computations that cannot be performed by ancient neural functions for numerosity. For example, the neural circuits responsible for numerosity cannot (on their own) represent -3 or $\sqrt{54}$, and yet this is simply represented in terms of public mathematical symbols (DeCruz 2008). This is because the symbolic representations are novel and unique. Initially, novelty results from the pressures of increasing social and economic complexity. Small roaming bands of foragers do not need to develop symbolic number systems; post-agricultural Neolithic societies settled in villages and towns do. A further issue is how novelty comes about from the ability to abstractly combine symbols and functions that apply to the symbols. I don't propose to try to answer that question here; however, we might think of this as a curiosity- and creativity-driven processes. Given uniquely human behavioural and neural plasticity and socio-cultural complexity we might expect an increasing drive towards cognitive innovation. This has certainly been the story of recent cultural evolution in modern human societies.

This concludes the discussion of mathematical cognition as enculturation. Now I turn to the objections.

6 The incredible shrinking system

Why not just shrink the cognitive system to brain-based systems? Is there a way to bridge

³⁰ I will not address the issue of what discovery amounts to here and will remain neutral on whether discovery reveals a platonic mathematical system or simply the logical relations between concepts.

the impasse between moderate and strong embedding? One argument concerns whether it makes any difference to cognitive science to consider, for example, the manipulation of public symbols to be cognitive processes (Sprevak 2010). Ultimately, to give a decisive answer to that question we would need to change our conception of cognitive processes to on-going dynamical interactions with the environment that loop through brain, body, and environment. However, weak and moderate embedded approaches do not work with such a conception of cognitive process; they work with an input-process-output style sandwich model, where processes supervene on bodily states and processes. For them, there is no reason to accept strong embedding, and much of the discussion has been based around thought experiments or abstract definitions rather than concrete examples.

However, even on a scaffolded view of cognition we can't deny the difference-making role the manipulations of symbols make to the completion of cognitive tasks. Manipulating public symbols is unique; there is a difference between internalised strategies for completing mathematical tasks and strategies for manipulating mathematical inscriptions. Our cognitive capacities cannot cope with long sequences of complex symbols and operations on them. This is why we must learn strategies and methods for writing out proofs. Symbol manipulation makes a unique difference to our ability to complete mathematical tasks, and we cannot simply ignore their role. If we take the approach of CI, then mathematical cognition is constituted by these bouts of symbol manipulation, and we cannot simply shrink the system back to the brain. The case for a strongly embedded approach to mathematical cognition depends upon the novelty and uniqueness of mathematical practices and dual component transformations. Our evolutionary endowments of numerosity are not up to the task of exact symbolic arithmetic and mathematics. Without symbolic number systems and sequential algorithms there would be no mathematical innovation. Mathematical innovation includes representational novelty: negative numbers, square roots, zero, etc., but also novel functions: multiplication, division,

etc. Novelty comes about from the ability to abstractly combine symbols and functions that apply to the symbols.

Uniquely, symbols represent quantities discretely, but there is also the unique human capacity of manipulating symbols in public space. We learn to manipulate symbols in public space and we continue to do so when completing cognitive tasks.

The entire system of mathematics is not contained in a single brain. Symbol systems are public systems of representations and practices for their manipulation. Mathematical practices are part of the niche that we inherit—they are part of our cultural inheritance.

6.1 Impermanent scaffolds?

Another objection concerns the impermanence of the scaffolding required for mathematical cognition. Once we have internalised the scaffolding of symbolic number systems, we have no further need for it, except for communication purposes. This claim would be proven if we did not continue to manipulate numerals when completing cognitive tasks. Even if we think that transformation only results in new internal representational resources, and that this just amounts to moderate embedding/scaffolding, we must also concede that most mathematics is conducted on the page.

Scaffolding theorists, like Sterelny, can endorse this idea; indeed they can agree with the bulk of the framework provided by CI whilst avoiding the constitutive claim. What they cannot do is deny that mathematical practice and the manipulation of physically laid-out symbols on the page is a difference maker for mathematical cognition. If you remove it, the ability to complete mathematical tasks drops considerably. To do so is to fly in the face of the empirical evidence from psychology (Landy & Goldstone 2007) and cognitive neuroscience (Dehaene & Cohen 2007; Ansari 2012). Consequently, it is clear that cognitive practices transform our mathematical abilities, lending weight to the CI approach.

The case I have presented in this paper is that symbols are not simply impermanent scaffolds.

folds, they are permanent scaffolds. They become part of the architecture of cognition (and not simply through internalisation). Mastery of symbol systems results in changes to cortical circuitry, altering function and sensitivity to a new, public, representational system. However, it also results in new sensori-motor capacities for manipulating symbols in public space. The case can be made in terms of what a symbol system is:

A symbol is a physical mark (or trace), either in physical space, or as a digital trace. Symbol systems contain rules and practices for interpreting symbols, for combining them, and for ordering and manipulating them. A large body of often tacit practices for interpreting and manipulating symbols is acquired. Scaffolding is not simply an amodal symbol with an abstract designation that needs to be learnt (or mapped onto some innate symbol); scaffolding is also how the symbols are physically arranged, how symbols are pushed from one place to the next in a regular fashion. Finally, scaffolding is also how we use our own bodies, eyes, ears, and hands to create and manipulate symbols.

7 Conclusion

I have presented a case for CI as a process of enculturation, with mathematical cognition as an example of the process of enculturation at work. I began by laying out the 4E landscape and locating CI within it, relative to enactivism and EM. In particular I showed how CI shares the interactive stance of enactivism and the constitutive stance of EM, but how it also differs from these. The main difference between CI and enactivism is that CI does not equate life and mind in the way that enactivism does. The main difference between CI and EM is that CI takes cultural practices to play a central role in the assembly of cognitive systems, whereas EM does not.

I then went on to outline the central concepts required to make sense of enculturation. The CI framework embraces both evolutionary continuity and transformation of existing cognitive circuitry in development. Our modern minds are built on archaic precursors by slow

incremental changes. However, modern humans are behaviourally plastic and scaffolded learning drives functional changes in our plastic brains. The developmental change from the ANS to the DNS is an example of how learning-driven changes to cortical function result in new abilities, but this would not happen without the novelty and uniqueness of mathematical symbols and the practices for manipulating them.

I also countered two standard objections: impermanence and shrinkage. The defence of CI rested on the novelty and uniqueness of mathematical practices and symbols.

If the CI framework is on the right track, then human cognitive evolution has resulted in minds that are flexible and interactive. Furthermore, cultural evolution has resulted in written symbol systems and practices for manipulating symbols that can be acquired (in development) by minds like ours. The uniqueness of modern human minds lies in their capacity for transformation.

References

- Adams, A. & Aizawa, K. (2008). *Defending the bounds of cognition*. Oxford, UK: Blackwell.
- Andrews, K. (2012). *Do apes read minds?: Toward a new folk psychology*. Cambridge, MA: MIT Press.
- Ansari, D. (2008). Effects of development and enculturation on number representation in the brain. *Nature Reviews Neuroscience*, 9 (4), 278-291. [10.1038/nrn2334](https://doi.org/10.1038/nrn2334)
- (2012). Culture and education: New frontiers in brain plasticity. *Trends in Cognitive Sciences*, 16 (2). [10.1016/j.tics.2011.11.016](https://doi.org/10.1016/j.tics.2011.11.016)
- Barbaras, R. (2010). Life and exteriority: The problem of metabolism. In J. Stewart, O. Gapenne & E. Di Paolo (Eds.) *Enaction toward a new paradigm for cognitive science* (pp. 89-122). Cambridge, MA: MIT Press.
- Barkow, J. H., Cosmides, L. & Tooby, J. (1992). *The adapted mind: Evolutionary psychology and the generation of culture*. Oxford, UK: Oxford University Press.
- Beer, R. D. (2000). Dynamical approaches to cognitive science. *Trends in Cognitive Sciences*, 4 (3), 91-99. [10.1016/S1364-6613\(99\)01440-0](https://doi.org/10.1016/S1364-6613(99)01440-0)
- Boyd, R. & Richerson, P. J. (2005). *The origin and evolution of cultures*. Oxford, UK: Oxford University Press.
- Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. Oxford, UK: Oxford University Press.
- (2011). Finding the mind. *Philosophical Studies*, 152 (3), 447-461. [10.1007/s11098-010-9598-9](https://doi.org/10.1007/s11098-010-9598-9)
- (2015). Embodied prediction. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Davies, S. (2012). *The artful species: Aesthetics, art, and evolution*. Oxford, UK: Oxford University Press.
- De Cruz, H. (2008). An extended mind perspective on natural numberrepresentation. *Philosophical Psychology*, 21 (4), 475-490. [10.1080/09515080802285289](https://doi.org/10.1080/09515080802285289)
- Dehaene, S. (1997). *The number sense: How the mind creates mathematics*. London, UK: Penguin.
- (2001). Précis of the number sense. *Mind & Language*, 16 (1), 16-36.
- (2009). *Reading in the brain: The new science of how we read*. London, UK: Penguin.
- Dehaene, S. & Cohen, L. (2007). Cultural recycling of cortical maps. *Neuron*, 56 (2), 384-398. [10.1016/j.neuron.2007.10.004](https://doi.org/10.1016/j.neuron.2007.10.004)
- (2011). The unique role of the visual word form area in reading. *Trends in Cognitive Sciences*, 15 (6), 254-262. [10.1016/j.tics.2011.04.003](https://doi.org/10.1016/j.tics.2011.04.003)
- Dehaene, S., Spelke, E., Pinel, P., Stanescu, R. & Tsivkin, S. (1999). Sources of mathematical thinking: Behavioral and brain-imaging evidence. *Science*, 284 (5416), 970-974. [10.1126/science.284.5416.970](https://doi.org/10.1126/science.284.5416.970)
- Dutilh Novaes, C. (2013). Mathematical reasoning and external symbolic systems. *Logique & Analyse*, 56 (221), 45-65.
- Finlayson, C. (2009). *The humans who went extinct: Why Neanderthals died out and we survived*. Oxford, UK: Oxford University Press.
- Gallagher, S. (2005). *How the body shapes the mind*. Oxford, UK: Oxford University Press.
- Gallese, V. (2008). Mirror neurons and the social nature of language: The neural exploitation hypothesis. *Social Neuroscience*, 3 (3-4), 317-333. [10.1080/17470910701563608](https://doi.org/10.1080/17470910701563608)
- Glenberg, A. (2010). Embodiment as a unifying perspective for psychology. *Cognitive Science*, 1 (4), 586-596. [10.1002/wcs.55](https://doi.org/10.1002/wcs.55)
- Godfrey-Smith, P. (1996). *Complexity and the function of mind in nature*. Cambridge, UK: Cambridge University Press.
- Hoffman, M. (2014). Evolution of the human brain: When bigger is better. *Frontiers in Neuroanatomy*, 8 (1). [10.3389/fnana.2014.00015](https://doi.org/10.3389/fnana.2014.00015)
- Huebner, B. (2013). *Macro cognition: Distributed minds and collective intentionality*. New York, NY: Oxford University Press.
- Hurley, S. (2010). The varieties of externalism. In R. Menary (Ed.) *The extended mind* (pp. 101-154). Cambridge, MA: MIT Press.
- Hutchins, E. (1995). *Cognition in the wild*. Cambridge, MA: MIT Press.
- (2011). Enculturating the supersized mind. *Philosophical Studies*, 152 (3), 437-446. [10.1007/s11098-010-9599-8](https://doi.org/10.1007/s11098-010-9599-8)
- Huttenlocher, P. R. (2002). *Neural plasticity: The effects of environment on the development of the cerebral cortex*. Cambridge, MA: Harvard University Press.
- Hutto, D. (2008). *Folk psychological narratives: The sociocultural basis of understanding reasons*. Cambridge, MA: MIT Press.
- Hutto, D. D. & Myin, E. (2013). *Radicalizing enactivism: Basic minds without content*. Cambridge, MA: MIT Press.
- Jefferies, B. (2010). The co-evolution of tools and minds: Cognition and material culture in the hominin lineage. *Phenomenology and the Cognitive Sciences*, 9 (4), 503-520. [10.1007/s11097-010-9176-9](https://doi.org/10.1007/s11097-010-9176-9)

- Laland, K. N., Odling-Smee, J. & Feldman, M. W. (2000). Niche construction, biological evolution and cultural change. *Behavioral and Brain Sciences*, 23 (1), 131-146. [10.1017/S0140525X00002417](https://doi.org/10.1017/S0140525X00002417)
- Landy, D., Allen, C. & Zednik, C. (2014). A perceptual account of symbolic reasoning. *Frontiers in Psychology*, 5 (275). [10.3389/fpsyg.2014.00275](https://doi.org/10.3389/fpsyg.2014.00275)
- Landy, D. & Goldstone, R. L. (2007). How abstract is symbolic thought? *Journal of Experimental Psychology*, 33 (4), 720-733. [10.1037/0278-7393.33.4.720](https://doi.org/10.1037/0278-7393.33.4.720)
- Lycett, S. J. & Gowlett, J. A. J. (2008). On questions surrounding the Acheulean “tradition”. *World Archaeology*, 40 (3), 295-315. [10.1080/00438240802260970](https://doi.org/10.1080/00438240802260970)
- Lyons, I. M., Ansari, D. & Beilock, S. L. (2012). Symbolic estrangement: Evidence against a strong association between numerical symbols and the quantities they represent. *Journal of Experimental Psychology: General*, 141 (4), 635-641. [10.1037/a0027248](https://doi.org/10.1037/a0027248)
- MacLean, E. L., Matthews, L. J., Hare, B. A., Nunn, C. L., Anderson, R. C., Aureli, F., Brannon, E. M., Call, J., Drea, C. M., Emery, N. J., Haun, D. B., Herrmann, E., Jacobs, L. F., Platt, M. L., Rosati, A. G., Sandel, A. A., Schroepfer, K. K., Seed, A. M., Tan, J., van Schaik, C. P. & Wobber, V. (2012). How does cognition evolve? Phylogenetic comparative psychology. *Animal cognition*, 15 (2), 223-238. [10.1007/s10071-011-0448-8](https://doi.org/10.1007/s10071-011-0448-8)
- Menary, R. (2007). *Cognitive integration: Mind and cognition unbounded*. London, UK: Palgrave Macmillan.
- (2010). Dimensions of mind. *Phenomenology and the Cognitive Sciences*, 9, 561-578. [10.1007/s11097-010-9186-7](https://doi.org/10.1007/s11097-010-9186-7)
- (2011). Our glassy essence: The fallible self in pragmatist thought. In S. Gallagher (Ed.) *The Oxford Handbook of the Self* (pp. 609-632). Oxford, UK: Oxford University Press.
- (2012). Cognitive practices and cognitive character. *Philosophical Explorations: An International Journal for the Philosophy of Mind and Action*, 15 (2), 147-164. [10.1080/13869795.2012.677851](https://doi.org/10.1080/13869795.2012.677851)
- (2014). Neuronal recycling, neural plasticity and niche construction. *Mind and Language*, 29 (3), 286-303. [10.1111/mila.12051](https://doi.org/10.1111/mila.12051)
- Menary, R. & Kirchhoff, M. (2014). Cognitive transformations and extended expertise. *Educational Philosophy and Theory*, 46 (6), 610-623. [10.1080/00131857.2013.779209](https://doi.org/10.1080/00131857.2013.779209)
- Nieder, A. & Dehaene, S. (2009). Representation of number in the brain. *Annual Review of Neuroscience*, 32, 185-208. [10.1146/annurev.neuro.051508.135550](https://doi.org/10.1146/annurev.neuro.051508.135550)
- Nieder, A., Diester, I. & Tudusciuc, O. (2006). Temporal and spatial enumeration processes in the primate parietal cortex. *Science*, 313 (5792), 1432-1435. [10.1126/science.1130308](https://doi.org/10.1126/science.1130308)
- Odling-Smee, F. J., Laland, K. N. & Feldman, M. F. (2003). Niche construction: The neglected process in evolution. *Monographs in Population Biology*, 37
- Price, C. J. & Devlin, J. T. (2011). The interactive account of ventral occipitotemporal contributions to reading. *Trends in Cognitive Sciences*, 15 (6), 246-253. [10.1016/j.tics.2011.04.001](https://doi.org/10.1016/j.tics.2011.04.001)
- Roepstorff, A., Niewöhner, J. & Beck, S. (2010). Enculturing brains through patterned practices. *Neural Networks*, 23 (8-9), 1051-1059.
- Rouselle, L. & Noël, M. P. (2008). The development of automatic numerosity processes in preschoolers: Evidence for numerosity-perceptual interference. *Developmental Psychology*, 44 (2), 544-560. [10.1037/0012-1649.44.2.544](https://doi.org/10.1037/0012-1649.44.2.544)
- Rowlands, M. (2010). *The new science of the mind: From extended mind to embodied phenomenology*. Cambridge, MA: MIT Press.
- Rupert, R. (2009). *Cognitive systems and the extended mind*. Oxford, UK: Oxford University Press.
- Semaw, S., Rogers, M. J., Quade, J., Renne, P. R., Butler, R. F., Dominguez-Rodrigo, M., Stout, D., Hart, W. S., Pickering, D. & Simpons, S. W. (2003). 2.6-Million-year-old stone tools and associated bones from OGS-6 and OGS-7, Gona, Afar, Ethiopia. *Journal of Human Evolution*, 45 (2), 169-177. [10.1016/S0047-2484\(03\)00093-9](https://doi.org/10.1016/S0047-2484(03)00093-9)
- Shultz, S., Nelson, E. & Dunbar, R. I. M. (2012). Hominin cognitive evolution: Identifying patterns and processes in the fossil and archaeological record. *Philosophical Transactions of the Royal Society*, 367 (1599), 2130-2140. [10.1098/rstb.2012.0115](https://doi.org/10.1098/rstb.2012.0115)
- Smart, J. J. C. (1959). Sensations and brain processes. *Philosophical Review*, 68, 141-156.
- Sprevak, M. (2010). Inference to the hypothesis of extended cognition. *Studies in History and Philosophy of Science Part A*, 41 (4), 353-362. [10.1016/j.shpsa.2010.10.010](https://doi.org/10.1016/j.shpsa.2010.10.010)
- Sterelny, K. (2003). *Thought in a hostile world : The evolution of human cognition*. Oxford, UK: Blackwell.
- (2010). Minds: Extended or scaffolded? *Phenomenology and the Cognitive Sciences*, 9 (4), 465-481. [10.1007/s11097-010-9174-y](https://doi.org/10.1007/s11097-010-9174-y)
- (2012). *The evolved apprentice : How evolution made humans unique*. Cambridge, MA: MIT Press.

- Stewart, J. R., Gapenne, O. & Di Paolo, E. A. (2010). *Enaction: Toward a new paradigm for cognitive science*. Cambridge, MA: MIT Press.
- Stich, S. (1983). *From folk psychology to cognitive science: The case against belief*. Cambridge, MA: MIT Press.
- Stout, D., Toth, N., Schick, K. & Chaminade, T. (2008). Neural correlates of Early Stone Age toolmaking: Technology, language and cognition in human evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363 (1499), 1939-1949. [10.1098/rstb.2008.0001](https://doi.org/10.1098/rstb.2008.0001)
- Thatcher, R. W. (1991). Maturation of the human frontal lobes: Physiological evidence for staging. *Developmental Neuropsychology*, 7 (3), 397-419. [10.1080/87565649109540500](https://doi.org/10.1080/87565649109540500)
- Theiner, G. (2013). Onwards and upwards with the extended mind: From individual to collective epistemic action. In L. Caporael, J. Griesemer & W. Wimsatt (Eds.) *Developing scaffolds* (pp. 191-208). Cambridge, MA: MIT Press.
- Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind*. Cambridge, MA: Harvard University Press.
- Toth, N. & Schick, K. D. (2006). *The Oldowan: Case studies into the earliest stone age*. Bloomington, IN: Stone Age Institute Press.
- Turner, J. S. (2000). *The extended organism: The physiology of animal-built structures*. Cambridge, MA: Harvard University Press.
- Uller, C., Jaeger, R., Guidry, G. & Martin, C. (2003). Salamanders (*Plethodon cinereus*) go for more: Rudiments of number in an amphibian. *Animal Cognition*, 6 (2), 105-112. [10.1007/s10071-003-0167-x](https://doi.org/10.1007/s10071-003-0167-x)
- Varela, F., Thompson, E. & Rosch, E. (1991). *The embodied mind*. Cambridge, MA: MIT Press.
- Vygotsky, L. (1981). The instrumental method in psychology. In J. Wertsch (Ed.) *The concept of activity in soviet psychology*. Armonk, NY: Sharpe.
- Wheeler, M. (2005). *Reconstructing the cognitive world: The next step*. Cambridge, MA: MIT Press.
- Wheeler, M. & Clark, A. (2008). Culture, embodiment and genes: Unravelling the triplehelix. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363 (1509), 3563-3575. [10.1098/rstb.2008.0135](https://doi.org/10.1098/rstb.2008.0135)
- Whiten, A., Hinde, R. A., Laland, K. N. & Stringer, C. B. (2011). Culture evolves. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366 (1567), 938-948. [10.1098/rstb.2010.0372](https://doi.org/10.1098/rstb.2010.0372)

Enriching the Notion of Enculturation: Cognitive Integration, Predictive Processing, and the Case of Reading Acquisition

A Commentary on Richard Menary

Regina E. Fabry

Many human cognitive capacities are rendered possible by enculturation in combination with specific neuronal and bodily dispositions. Acknowledgment of this is of vital importance for a better understanding of the conditions under which sophisticated cognitive processing routines could have emerged on both phylogenetic and ontogenetic timescales. Subscribing to enculturation as a guiding principle for the development of genuinely human cognitive capacities means providing a description of the socio-culturally developed surrounding conditions and the profound neuronal and bodily changes occurring as a result of an individual's ongoing interaction with its cognitive niche. In this commentary, I suggest that the predictive processing framework can refine and enrich important assumptions made by the theory of cognitive integration and the associated approach to enculturated cognition. I will justify this suggestion by considering several aspects that support the complementarity of these two frameworks on conceptual grounds. The result will be a new integrative framework which I call enculturated predictive processing. Further, I will supplement Richard Menary's enculturated approach to mathematical cognition with an account of reading acquisition from this new perspective. In sum, I argue in this paper that the cognitive integrationist approach to enculturated cognition needs to be combined with a predictive processing style description in order to provide a full account of the neuronal, bodily, and environmental components giving rise to cognitive practices. In addition, I submit that the enculturated predictive processing approach arrives at a conceptually coherent and empirically plausible description of reading acquisition.

Keywords

Cognitive integration | Cognitive transformation | Enculturation | Neural plasticity | Neuronal reuse | Predictive processing | Reading acquisition | Scaffolded learning

1 Introduction

In his target paper *Mathematical Cognition: A Case of Enculturation*, Richard Menary investigates the conditions under which phylogenetically recent, socio-culturally shaped target phenomena within cognitive science such as mathematics, reading, and writing have emerged.

Resting on his theory of cognitive integration (CI; e.g., [Menary 2007a](#)), he starts from the idea that these processes are fully continuous with phylogenetically older ones (*evolutionary continuity*). This type of continuity is justified by the assumption that the evolution of neur-

Commentator

[Regina E. Fabry](#)

fabry@students.uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Target Author

[Richard Menary](#)

richard.menary@mq.edu.au

Macquarie University
Sydney, Australia

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

onal reuse mechanisms allows for the redeployment of cortical circuits for phylogenetically recent functions (Anderson 2010; Anderson & Finlay 2014). Ontogenetically, neuronal reuse is a precondition of *learning driven plasticity* (LDP), which “can result in both structural and functional changes in the brain” (Menary this collection, p. 8). That is, the human brain is assumed to be neuronally plastic so that its processing routines are altered as the individual acquires new cognitive abilities (Ansari 2012). However, the acquisition of new cognitive abilities takes place within

[...] a highly structured cognitive niche that contains not only physical artefacts, but also: representational systems that embody knowledge (writing systems, number systems, etc.); skills and methods for training and teaching new skills (Menary & Kirchhoff 2014); practices for manipulating tools and representations. (Menary this collection, p. 6)

It is this cognitive niche that provides the resources for *scaffolded learning*, which allows the individual to acquire new cognitive abilities through its ongoing embodied interaction with its socio-cultural environment. Together, LDP and scaffolded learning lead to cognitive transformations that augment the individual’s cognitive capacities through ontogenesis: “Cognitive transformations result from our evolved plasticity and scaffolded learning in the developmental niche” (Menary this collection, p. 8).¹ The result of cognitive transformation is the acquisition of a sufficient degree of expertise in performing a certain *cognitive practice*. Cognitive practices are normatively constrained to the extent that socio-culturally shaped procedures work in close interaction with the cognitive niche: They “[...] are culturally endowed (bodily) manipulations of informational structures” (Menary this collection, p. 4), such as manipu-

lations of tokens of a representational writing system, and they serve to complete a cognitive task. In order to describe the transformational processes by which cognitive practices are acquired, Menary introduces the notion of *enculturation*: “Enculturation rests on the acquisition of cultural practices that are cognitive in nature” (*ibid.*). That is, enculturation refers to any cognitive transformation that is rendered possible by LDP and the individual’s ongoing interaction with its cognitive niche. As a proof of concept, Menary (this collection) deals with mathematical cognition and describes the ways in which individuals acquire expertise in manipulating a public, socio-culturally developed mathematical symbol system. Relying on a set of empirical results, he arrives at the conclusion that precise mathematical operations are rendered possible by the recruitment of a neuronal sub-system during ontogeny. In contrast to the evolved approximate number system (ANS), which allows for subitizing and is also present in other animals, the neuronal realization of the discrete number system (DNS) heavily depends on LDP, the individual’s immersion into its cognitive niche, and its active participation in scaffolded learning routines. Thus, the acquisition of mathematical skills is an important example of enculturation.

The purpose of this commentary is to enrich and refine the enculturated approach. First, I will propose that the predictive processing framework provides conceptual and explanatory tools for describing and explaining the neuronal and extracranial bodily mechanisms underlying cognitive practices and enculturation. Thus, I will accept the challenge to combine “[...] the dynamical nature of causal commerce between world, body, and brain and the inferential free energy principle that allows their unification in one account” (Hohwy this collection, p. 18). I will argue that a new integrative framework that views CI and predictive processing as complementary is able to meet this challenge. Second, I will illustrate this by presenting reading acquisition as a paradigmatic case of enculturated cognition. In particular, I will demonstrate that a position that combines the enculturated approach with predictive processing,

¹ More precisely, according to Menary (2014, p. 293) it is scaffolded learning that renders LDP possible in the course of cognitive development of individuals: “Both structural and functional plasticity can result from both endogenous and exogenous sources, but here the focus is on structural and functional changes driven by scaffolded learning.”

which I call enculturated predictive processing, leads to a parsimonious and conceptually coherent account of reading acquisition that helps interpret and unify a vast array of recent empirical findings.

2 Towards a more complete approach to enculturation: Cognitive integration and predictive processing

In order to appreciate the descriptive power of the enculturated approach, it is necessary to specify the mechanistic underpinnings of the acquisition of cognitive practices. In his summary of the CI framework, Menary ([this collection](#), p. 2) argues that “[a]lthough the framework is unified by a dynamical systems description of the evolution of processing in the hybrid and multi-layered system, it recognises the novel contributions of the distinct processing profiles of the brain, body and environment.” However, the dynamical systems style approach to the acquisition and enactment of cognitive practices in the version first introduced in Menary (2007a, pp. 42-48) does not exhaustively specify the distinct, yet highly interactive neuronal and bodily components of cognitive processing. Furthermore, it does not account for LDP, simply because it remains neutral to the concrete realization of its neuronal component system. Finally, the dynamical systems approach, on Menary’s construal, helps illustrate what the interactive contribution of neuronal and extracranial bodily components to human cognition might amount to. Yet, it does not spell out the mutual influence that neuronal and extracranial bodily components have over each other.

This is where predictive processing (PP) enters the picture. In the remainder of this commentary I will argue that the PP approach provides the resources for a more detailed account of how human cognitive systems become enculturated and how they are subject to integrated cognition.

2.1 Cognitive integration: Five theses about human cognition

In its original version (cf. Menary 2007a), CI is constituted by five theses. They emphasize the

different aspects that are crucial for an integrationist approach to cognitive processing: 1. Human cognition is continuous with animal cognition on both diachronic and synchronic scales. However, it has a special status in that it is situated in a particular cognitive niche and heavily rests upon neural plasticity which is itself an adaptation (*continuity thesis*). 2. Certain cognitive processes are hybrid because they are constituted by neuronal and extracranial bodily components (*hybrid mind thesis*). 3. In the course of ontogenetic hybrid cognitive processing, both the constitutive neuronal and extracranial bodily functions are transformed (*transformation thesis*). 4. The bodily manipulation of specific environmental resources plays a crucial functional role in integrated cognitive processes (*manipulation thesis*). 5. These manipulations are constrained by cognitive norms, which are acquired through learning, and which realize socio-culturally developed habits for the interaction with cognitive resources (*cognitive norms thesis*).

In addition to the continuity thesis and the cognitive transformation thesis, which are given centre stage in Menary’s target paper, the hybrid mind thesis is important in that it acknowledges the close interaction of neuronal and extra-neuronal bodily sub-processes in the completion of cognitive tasks. In other words, certain cognitive processes “involve the integration of neural manipulations of vehicles and bodily manipulations of environmental vehicles” (Menary 2010, p. 236; see also Menary 2007b, p. 627). The notion of bodily manipulation as it is used here goes back to Mark Rowlands’ (1999, pp. 23f) account of *environmentalism*, which claims that “cognitive processes are, in part, made up of manipulation of relevant structures in the cognizer’s environment”. In this context, manipulation is defined as “any form of bodily interaction with the environment – manual or not, intrusive or otherwise – which makes use of the environment in order to accomplish a given task” (*ibid.*, p. 23). Thus, subscribing to the manipulation thesis amounts to the assumption that “[c]ognitive processing often involves these online bodily manipulations of the cognitive niche, sometimes as individuals and sometimes

in collaboration with others” (Menary [this collection](#), p. 3). Importantly, it is assumed that extracranial bodily manipulations causally interact with neural sub-processes, thereby stressing the hybridity of cognitive processes (cf. Menary 2007a, p. 138). In addition to highlighting the constitutive role of embodied engagements with “external” cognitive resources as proposed by Rowlands (1999), cognitive integrationists claim that the manipulation of these resources is constrained by cognitive norms. In this vein, Menary (2007a, p. 5; 2010, p. 233) argues that “[o]ur abilities to manipulate the extrabodily environment are normative and are largely dependent on our learning and training histories.” The idea that certain cognitive abilities are normatively structured thus concerns the individual’s interaction with specific resources provided by the cognitive niche. Importantly, the normatively constrained ways in which environmental resources are integrated into cognitive processes are shared by many individuals. Put differently, the normativity of cognitive practices helps “[...] stabilise and govern interactive thought across a population of similar phenotypes” (Menary [this collection](#), p. 4). Furthermore, the acquisition of a certain cognitive practice is tightly connected with the acquisition of the relevant cognitive norms in the course of scaffolded learning. This is because “we learn cognitive practices by learning the cognitive norms that govern the manipulation of vehicles” (Menary 2007b, p. 628).

From these five theses defended by CI it follows that there should be two distinct, yet interdependent levels of description for cognitive practices. First, there is the social level of description. On this level, cognitive practices need to be approached by highlighting the interactive, cooperative cognitive achievements of a large group of individuals sharing the same cognitive niche. Second, cognitive practices can be investigated by approaching them on an individual level of description. In this case, the acquisition and enactment of a certain cognitive practice is described with regards to a certain individual. However, any individual level description needs to acknowledge that certain cognitive capacities of an enculturated individual

are rendered possible only by the individual’s ongoing interaction with its socio-culturally shaped environment in normatively constrained ways. This means to do justice to the broader socio-cultural context of enculturated cognition, while being interested in a precise description of its neuronal and extracranial bodily sub-components. In this commentary I will operate on the individual level of description without denying that it is important to develop a fine-grained description on the social level by specifying the properties of a certain cognitive niche and the conditions under which it could have emerged.

To this end, I will now proceed by summarizing the most important features of the predictive processing (PP) approach that will help specify the mechanistic underpinnings of enculturated cognition.

2.2 An outline of predictive processing

Recently, the idea that human perception, action, and cognition can be described and explained in terms of hierarchically organized predictive processing mechanisms implemented in the human brain has enjoyed widespread attention within cognitive neuroscience (e.g., Friston 2005, 2010; Friston et al. 2012), philosophy of mind, and philosophy of cognitive science (e.g., Clark 2012, 2013, [this collection](#); Hohwy 2011, 2012, 2013, 2014, [this collection](#); Seth [this collection](#)). The overall epistemic goal of this emerging approach is to describe perceptual, sensorimotor, and cognitive target phenomena within a single framework by relying on unifying mechanistic principles. Accounts of PP generally assume that human perception, action, and cognition are realized by Bayesian probabilistic generative models implemented in the human brain. Since the human brain does not have immediate access to the environmental causes of sensory effects, it has to infer the most probable state of affairs in the environment giving rise to sensory data (cf. Seth [this collection](#), pp. 4f). PP approaches solve this *inverse problem* by assuming that generative models in accordance with Bayes’ rule are implemented in the human brain. On this construal, a generative model

“[...] aims to capture the statistical structure of some set of observed inputs by tracking [...] the causal matrix responsible for that very structure” (Clark 2013, p. 182). In order to be able to infer the causes of sensory effects, generative models encode probability distributions. Each generative model provides several hypotheses about the causes of a certain sensory input. The system has somehow to ‘decide’ which hypothesis needs to be chosen in order to account for the cause of the sensory effect. The descriptive power of Bayes’ rule lies in its capacity to capture the probabilistic estimations underlying these choices. Applied to the case of human perception, action, and cognition, Bayesian generative models are assumed to be realized in hierarchically organized structures comprising multiple, highly interactive low- and high-level cortical areas. This is referred to as the *Bayesian brain hypothesis* (cf. Friston 2010, p. 129). The hierarchical organization of probabilistic generative models is combined with a specific version of *predictive coding*, where predictive coding “depicts the top-down flow as attempting to predict and fully ‘explain away’ the driving sensory signal, leaving only any residual ‘prediction errors’ to propagate forward within the system” (Clark 2013, p. 182). That is to say, selected hypotheses inform prior predictions about the sensory input to be expected at each level of the hierarchy. These predictions fulfil the function of encoding knowledge about statistical regularities of patterns in the observable (or any imaginable) world. This hypothesis selection proceeds in accordance with Bayes’ rule. The processing of sensory input gives rise to prediction errors. Prediction errors carry neuronally realized information about “[...] residual differences, at every level and stage of processing, between the actual current signal and the predicted one” (Clark this collection, p. 4). Importantly, it is only prediction errors, and not sensory input *per se*, that are fed forward within the hierarchy (cf. Clark 2013, pp. 182f; Hohwy 2012, p. 3, 2013, p. 47, 2014, p. 4). The overall aim of this multi-level processing mechanism is to *minimize prediction error*, that is, to reduce or to ‘explain away’ the discrepancy between predictions and the actually given sensory input

that is an effect of environmental (or bodily) causes (cf. Clark 2013, p. 187; Hohwy 2011, p. 269, 2013, p. 88). This is known as *prediction error minimization*.²

Prediction error minimization is a special way of minimizing *free energy* in accordance with the principle “that any self-organizing system that is at equilibrium with its environment must minimize its free energy” (Friston 2010, p. 127). Applied to human perception, cognition, and action, minimizing free energy means minimizing the amount of unbound energy available to the perceiving, cognizing, and acting organism. This is where prediction error enters the picture. As Andy Clark (2013, p. 186) puts it, “[p]rediction error reports this information-theoretic free energy, which is mathematically constructed so as always to be greater than ‘surprisal’ (where this names the sub-personally computed implausibility of some sensory state given a model of the world [...]).” The relationship between free energy and surprisal then is that “[...] free energy is an upper bound on surprise, which means that if agents minimize free energy, they implicitly minimize surprise” (Friston 2010, p. 128). Suprisal, however, cannot be estimated directly by the system, because “there is an infinite number of ways in which the organism could seek to minimize surprise and it would be impossibly expensive to try them out” (Hohwy 2012, p. 3). The solution to this problem lies in implicitly minimizing surprisal (and its upper bound, i.e., free energy) by minimizing prediction error (cf. Hohwy 2013, p. 85, this collection, 3; see also Seth this collection, p. 6). It is exactly here where prediction

2 On a neuronal level of description, hierarchical generative models are assumed to be neuronally realized by multiple connections across low- and high-level cortical areas. Each level within the cortical hierarchy is connected to the next subordinate and supraordinate level, thereby ensuring effective inter-level message passing (cf. Hohwy 2013, pp. 67f). According to Clark (2013, p. 187), predictive generative models are implemented in “a kind of duplex architecture”. This means that there are distinct neuronal units dedicated to the representation of predictions of environmental (or bodily) causes, so-called *representation units*, on the one hand, and those dedicated to the encoding of prediction error, so-called *error units*, on the other (cf. *ibid.*; Friston 2005, p. 829). To date, a detailed account of the concrete neuronal realization of these functionally distinct units of message-passing is still missing (cf. *ibid.*). However, it is hypothesized that representation units might correspond to superficial pyramidal cells, while error units might correspond to deep pyramidal cells (cf. Friston et al. 2012, p. 8; see also Clark 2013, pp. 187f).

error minimization avails itself as a tractable expression of more general life-sustaining mechanisms.

Prediction error minimization can be achieved in two distinct, yet complementary ways. The first of these is *perceptual inference*, which can be described as

[...] an iterative step-wise procedure where a hypothesis is chosen, and predictions are made, and then the hypothesis is revised in light of the prediction error, before new and hopefully better predictions are made on the basis of the revised hypothesis. (Hohwy 2013, p. 45)

That is, prediction errors are propagated up the hierarchy leading to an adjustment of the initial hypothesis, thereby achieving an approximation of the hypothesis generating the predictions and the actually given input. The adjustment of predictions and hypotheses in the face of feed-forward prediction error occurs at every level of the hierarchy until any prediction error is accommodated. This complex process comprising multiple levels is known as perception: “Perception thus involves ‘explaining away’ the driving (incoming) sensory signal by matching it with a cascade of predictions pitched at a variety of spatial and temporal scales” (Clark 2013, p. 187; see also Clark 2012, p. 762).

On Andy Clark’s account of PP, one important consequence of this is that the traditional distinction between perception and cognition becomes blurred. It is replaced by a reconceptualization of perceptual and cognitive processes as a continuous employment of the same prediction error minimizing mechanism on multiple scales:

All this makes the lines between perception and cognition fuzzy, perhaps even vanishing. In place of any real distinction between perception and belief we now get variable differences in the mixture of top-down and bottom-up influence, and differences of temporal and spatial scale in the internal models that are making predictions. Top-level (more ‘cognitive’) models

intuitively correspond to increasingly abstract conceptions of the world, and these tend to capture or depend upon regularities at larger temporal and spatial scales. Lower-level (more ‘perceptual’) ones capture or depend upon the kinds of scale and detail most strongly associated with specific kinds of perceptual contact. (Clark 2013, p. 190)

Consequently, processes typically associated with perception or cognition can only be distinguished by considering the temporal and spatial resolution of the instantiation of PP mechanisms and the levels at which model revision ensues, respectively. This relationship between perception and cognition becomes important once we consider how enculturated cognition has been rendered possible on both phylogenetic and ontogenetic time scales. For it helps specify how evolutionary continuity could have been rendered possible in the first place. The evolutionary development of perception and cognition (and, as we shall see, of action too) may have proceeded from more perceptual generative models present in many other animals to more cognitive generative models exclusively realized in humans. This is in line with Roepstorff’s (2013, p. 45) observation that “[t]he underlying neural models are basically species-unspecific, and the empirical cases move back and forth between many different model systems.” Referring to this observation, Clark (this collection, p. 14) emphasizes that “[t]he basic elements of the predictive processing story, as Roepstorff (2013, p. 45) correctly notes, may be found in many types of organism and model-system.” Thus, while certain (lower-level) model parameters and processing stages of prediction error minimization are shared by many organisms, there certainly are specific (higher-level) processing routines that are shared only by enculturated human organisms in a certain cognitive niche.

Furthermore, the idea that perception and cognition are continuous is relevant for considerations of the ontogenetic development of enculturated cognitive functions. This is because it anchors higher-order cognitive operations in

more basic perceptual processes and thus allows for a fine-grained description of a certain developmental trajectory leading to cognitive transformation. Bearing in mind the hierarchical structure of generative models, another interesting consequence of the PP style approach to perception and cognition is that lower (i.e., more perceptual) levels of the generative model influence higher (i.e., more cognitive) levels by means of fed-forward prediction error. Vice versa, higher levels of the hierarchical generative model influence lower levels by means of fed-backward predictions (cf. Hohwy 2013, p. 73). This will become more important when we explore how reading acquisition can be described as an ongoing enculturating process of prediction error minimization.

Perceptual inference is only one way of minimizing prediction error. The second is *active inference*, where “[...] the agent will selectively sample the sensory input it expects” (Friston 2010, p. 129). The idea is that the system can minimize prediction error by bringing about the states of affairs (i.e., the environmental hidden causes) that are predicted by a certain hypothesis. This is achieved by performing any type of bodily movements, including eye movements, that make the selected prediction come true. The predictions at play in active inference are *counterfactual*, because

[...] they say how sensory input *would* change if the system *were* to act in a certain way. Given that things are not actually that way, prediction error is induced, which can be minimized by acting in the prescribed way. (Hohwy 2013, p. 82; italics in original; see also Clark this collection, p. 6; Friston et al. 2012, p. 2)

Accordingly, in active inference the selected prediction is held constant and leads to bodily activities that minimize prediction error by altering the sensory input such that it confirms the prediction. Therefore, active inference is of crucial importance for prediction error minimization, “[...] since it provides the only way (once a good world model is in place and aptly activated) to actually alter the sensory signal so as

to reduce sensory prediction error” (Clark 2013, p. 202).

This suggests that perceptual and active inference, or perception and bodily action for that matter, mutually influence each other, thereby minimizing prediction errors and optimizing hypotheses generating ever new predictions. However, perceptual and active inference have a “different direction of fit” (Hohwy 2013, p. 178; see also Hohwy this collection, p. 13; Clark this collection, p. 7).³ This is because in perceptual inference, predictions are aligned to the sensory input, while active inference is a matter of aligning the sensory input to the predictions. It follows “[...] that to optimally engage in prediction error minimization, we need to engage in perceptual inference and active inference in a complementary manner” (Hohwy 2013, p. 91). Since both perceptual and active inference are aimed at minimizing prediction error and optimizing generative models, “[p]erception and action [...] emerge as two sides of a single computational coin” (Clark 2012, p. 760).

As emphasized earlier, perception and cognition are deeply related to the extent that both phenomena are the result of the same underlying functional and neuronal mechanisms. By extension, action is also deeply intertwined with cognition. This follows from the assumptions that 1. perception and cognition are continuous and 2. perception and action are subject to the same principles of prediction error minimization. As Seth (this collection, p. 5) puts it, both ways of prediction error minimization “[...] unfold continuously and simultaneously, underlining a deep continuity between perception and action [...]” Yet, perceptual and active inference fulfil distinct functional roles in their ongoing attempt to minimize prediction error. This becomes even more obvious once we take the free energy principle into account: “The free energy principle [...] does not posit any fundamental difference between perception and action. Both fall out of different reorganizations of the principle and come about mainly as different direc-

3 The notion of two functions having “a different direction of fit” originates in J. L. Austin’s (1953, p. 234) speech act theory and in G. E. M. Anscombe’s (1963, p. 56) example illustrating how words and states of affairs can relate to each other. I would like to thank Thomas Metzinger for pointing out the philosophical history of this notion.

tions of fit for prediction error minimization [...]” (Hohwy this collection, p. 13). Active inference plays a crucial role in cognition (understood as prediction error minimization comprising many higher-level predictions), for it helps minimize prediction error throughout the cortical hierarchy by bringing about the states of affairs in the environment that are predicted on higher levels. Therefore, on Clark’s (2013, p. 187) account, which he dubs *action-oriented predictive processing*, prediction error minimization “[...] depicts perception, cognition and action as profoundly unified and, in important respects, continuous.”

PP accounts of human perception, action, and cognition distinguish between first-order and second-order statistics. In contrast to first-order statistics, which amount to minimizing prediction error by means of perceptual and active inference, second-order statistics are concerned with estimating the *precision* of prediction error. In second-order statistics, the influence of feed-forward prediction error on higher levels of the hierarchical generative model is dependent upon its estimated precision. Neuronally, the estimation of precision is captured in terms of increasing or decreasing the *synaptic gain* of specific error units (cf. Feldman & Friston 2010, p. 2). That is, “[t]he more precision that is expected the more the gain on the prediction error in question, and the more it gets to influence hypothesis revision” (Hohwy 2013, p. 66; see also Friston 2010, p. 132). Conversely, if the precision is expected to be poor on the basis of second-order statistics, the synaptic gain on the error unit is inhibited such that the prediction on the supraordinate level is strengthened (cf. *ibid.*, p. 123). It has been proposed that precision estimation is equivalent to attention. This means that “attention is nothing but optimization of precision expectations in hierarchical predictive coding” (Hohwy 2013, p. 70; see also Feldman & Friston 2010, p. 2). For current purposes, it is sufficient to focus in the main on first-order statistics. However, it is important to bear in mind the crucial modulatory role precision estimation plays in prediction error minimization.

2.3 Combining cognitive integration and predictive processing

To what extent is it feasible to describe the mechanisms underlying cognitively integrated processes and enculturated cognition in terms of prediction error minimization? After having summarized CI and the core ideas of the PP framework I will argue in this section that there are many aspects of the CI approach that can be enriched by making a crucial assumption, namely that PP can account for many components constituting cognitive practices on at least functional and neuronal levels of description.

First, a major conceptual consequence of PP is that perception, action, and cognition are both continuous and unified, if this approach proves correct. This is because they follow the same principles of prediction error minimization, yet are characterized by important functional differences. This kind of complementarity fits neatly with the *hybrid mind thesis* defended by CI. Recall that the hybrid mind thesis claims that cognitive processes are constituted by both neuronal and extracranial bodily components. By taking prediction error minimization into account, this claim can be cashed out by assuming that the neuronal components are equal to perceptual inferences at multiple levels of the cortical hierarchy, while the bodily components are mechanistically realized by active inferences. The hybrid mind thesis emphasizes the indispensable, close and flexible coordination of neuronal and bodily components responsible for the completion of a cognitive task. The PP framework, or so I shall argue, provides the resources for a careful description of the underlying mechanisms at play. It does so by depicting human organisms as being constantly engaged in prediction error minimization by optimizing hypotheses in the course of perceptual inference and by changing the stimulus array in the course of active inference.

A second advantage of the prediction error minimization framework is that it helps cash out the *manipulation thesis*. This thesis, recall, states that “the manipulation of external vehicles [is] a prerequisite for higher cognition and embodied engagement [is] a precondition

for these manipulative abilities” (Menary 2010, p. 232). In terms of the PP framework, bodily manipulation can be understood as an instance of active inference occurring in specific contexts. That is, in order to complete a certain cognitive task, the system changes its sensory input by altering certain components of its cognitive niche. This becomes even more obvious once we take into account that embodied activity is also a means of increasing confidence in sensory input by optimizing its precision. As suggested by Hohwy (this collection, p. 6), “expected precision drives action such that sensory sampling is guided by hypotheses that the system expects will generate precise prediction error.” Applied to an organism’s interaction with its socio-culturally shaped environment, Hohwy (2013, p. 238) argues “[...] that many of the ways we interact with the world in technical and cultural aspects can be characterized by attempts to make the link between the sensory input and the causes more precise (or less uncertain).” However, bodily manipulation is more than just a contributing factor to prediction error minimization (and precision optimization). In order to acknowledge this, we need to take into account that bodily manipulations are a crucial component of the performance of cognitive practices. In the performance of a cognitive practice, the minimization of prediction error and the optimization of precision is not an end in itself. Rather, it serves to facilitate the completion of a certain cognitive task. Furthermore, the concrete bodily manipulations given in terms of active inference are subject to cognitive norms that constrain the ways in which human organisms interact with cultural resources, such as tokens of a representational writing system. That is to say that the performance of a cognitive practice is not an individualistic enterprise. Rather, in completing a cognitive task, the individual is deeply immersed into a socio-cultural context which is shared by many human organisms.

Third, it is the normative constraints on cognitive practices that render their performance efficient and, in many cases at least, successful. This is because compliance with these norms induces what Andy Clark (2013, p. 195)

calls “path-based idiosyncrasies”. That is, one of the reasons why the coordination of neuronal and bodily components in the manipulation of cultural resources is beneficial certainly is that it takes place in a normatively constrained “multi-generational development of stacked, complex ‘designer environments’ for thinking such as mathematics, reading, writing, structured discussion, and schooling” (ibid.). That is to say that the performance of cognitive practices in compliance with certain norms has the overall advantage of reducing cognitive effort, which can be captured as the minimization of overall prediction error and the optimization of precision on a sub-personal level of description. At the same time, however, cognitive practices themselves can be described, or so I shall argue, as having prediction error minimization as their underlying mechanism. This double role of cognitive practices, described in terms of prediction error minimization, can be fully appreciated once we consider the cognitive transformations brought about by the ongoing interaction with cultural resources.

Fourth, our cognitive capacities and the various ways we complete cognitive tasks are profoundly augmented by our neuronal and bodily engagements with the socio-culturally structured environment through ontogenesis (cf. Menary 2006, p. 341). Put differently, “cognitive transformations occur when the development of the cognitive capacities of an individual are sculpted by the cultural and social niche of that individual” (Menary this collection, p. 8). This niche includes mathematical symbol systems, representational writing systems, artifacts, and so forth. It is this immersion and, importantly, the scaffolding provided by other inhabitants of the cognitive niche that ideally lead to the transformation of neuronal and extracranial bodily components constituting cognitive processes, to enculturation that is. The PP framework, or so I shall argue, offers a highly promising account of learning that is most suitable for a sub-personal level description of cognitive transformation. On the construal of PP, learning flows naturally from the mechanism of prediction error minimization. For learning can generally be construed as a sub-personally real-

ized strategy of optimizing models and hypotheses in the face of ever new prediction error: “Learning is then viewed as the continual updating of internal model parameters on the basis of degree of predictive success: models are updated until they can predict enough of the signal” (Hohwy 2011, p. 268). Broadly understood, ‘learning’ thus figures as an umbrella term referring to the ongoing activity of prediction error minimization and model optimization throughout the lifetime of a human organism. This is because potentially ever new and “surprising” sensory signals need to be “explained away” by perceptual and active inference. For current purposes, however, “learning” can also be understood in a rather narrow sense as the acquisition of a certain skill, which is also subject to prediction error minimization through perception, action, cognition, and the modulation of attention. It is the individual’s socio-culturally structured environment that delivers new sensory signals helping optimize parameters of the generative model:

But those training signals are now delivered as part of a complex developmental web that gradually comes to include all the complex regularities embodied in the web of statistical relations among the symbols and other forms of socio-cultural scaffolding in which we are immersed. We thus self-construct a kind of rolling ‘cognitive niche’ able to induce the acquisition of generative models whose reach and depth far exceeds their apparent base in simple forms of sensory contact with the world. (Clark 2013, p. 195)

However, complex skills that are targeted at the completion of cognitive tasks cannot be learned simply by being exposed to the right kind of “training signal” in the cognitive niche. What is additionally needed is engagement in activities that are scaffolded by inhabitants of that cognitive niche who have already achieved a sufficient degree of expertise. This is what Menary (this collection) calls “scaffolded learning”. From the perspective of PP, this amounts to the strategy of exposing predictive systems to

highly structured, systematically ordered patterns of sensory input in the cognitive niche. This, however, needs to be complemented by a fine-grained personal-level description of the kind of interactions between experts and novices that is needed in order to pass on the right set of cognitive norms. Furthermore, the kind of cognitive transformation at play here requires a description of the neuronal changes that are correlated with the acquisition of a certain cognitive practice. That is, we need a more fine-grained account of LDP and how it might be realized in the human cortex. From the perspective of the PP framework, one plausible conjecture at this point is that LDP can be captured in terms of *effective connectivity*. Effective connectivity reports the causal interaction of neuronal assemblies across multiple levels of the cortical hierarchy (and across different brain areas) as a result of attention in terms of precision estimation. This line of reasoning is implied by Clark (2013, p. 190) who argues that “[a]ttention [...] is simply one means by which certain error-unit responses are given increased weight, hence becoming more apt to drive learning and plasticity, and to engage in compensatory action.” This last point is important, since it stresses that it is not only perceptual inference that drives learning and contributes to the improvement of generative models, but also active inference. However, this approach to the acquisition of action patterns in concert with an optimization of precision might raise the worry that learning is depicted here as being a rather internalistic, brain-bound affair. But once we acknowledge that it is the performance and ongoing improvement of embodied active inferences that play an indispensable functional role in the completion of cognitive tasks, it becomes obvious that this worry is not warranted. For it is the efficient interaction of neuronal and extracranial bodily components (i.e., perceptual and active inferences in terms of PP) that results from learning and the efficient engagement of human organisms with their environment. Furthermore, LDP can now be considered in terms of the precision-weighted optimization of hypotheses throughout the cortical hierarchy and the ever new patterns of effective con-

nectivity, as new cognitive practices are acquired and successfully performed. The sub-personal description of cognitive transformation in terms of prediction error minimization also does justice to neuronal reuse as a guiding principle of the allocation of neuronal resources for phylogenetically recent cognitive functions such as arithmetic or reading.

From this, the following question arises: What is the actual relationship between CI and PP supposed to be and what is the scope of this theory synthesis? First of all, the position developed in this commentary is neutral with regards to metaphysical consequences that may or may not result from the idea that CI and PP can be integrated into a unified theoretical framework. Rather, this position has an instrumentalist flavour to the extent that it tries to answer the question by which means socio-culturally shaped target phenomena can be best investigated both conceptually and empirically. Thus, the combination of CI and PP is valid only to the extent that it displays great descriptive as well as predictive power and is supported by many results stemming from empirical research. As such, the new approach on offer here is contingent upon the current state of research in cognitive science. It is falsifiable by new empirical evidence or convincing conceptual considerations that directly speak against it. Furthermore, it sidesteps the concern that PP and the underlying free energy principle might be trivial because they can be applied to any target phenomenon by telling a “just-so story”. This is because the combination of CI and PP is applied to specific domains, namely to classes of cognitive processes that count as cognitive practices, with reading being the paradigm example.⁴ Thus the approach advocated can be seen as a modest contribution to the project aiming at a “[...] translation into more precise, constricted applications to various domains, where predictions can be quantified and just-so stories avoided” (Hohwy [this collection](#), p. 14).

The idea that CI and PP can be combined can lead to different degrees of commitment.⁵

First, I do not assume that CI *necessarily requires* PP. Hypothetically, it is conceivable that another theory of neuronal and bodily functioning might be more suited to cashing out cognitive practices and enculturation more convincingly and more extensively. To date, PP appears to be the best unifying framework that helps specify exhaustively the functional and neuronal contributions of bodily and neuronal sub-processes giving rise to cognitive practices and enculturation. This is because PP offers a fine-grained functional and neuronal description of perception, action, cognition, attention, and learning that does justice to the complex interactions stipulated by CI and the associated approach to enculturation.

Second, it could be assumed that CI and PP are merely compatible. This would mean that CI and PP were self-sufficient and co-existent theoretical frameworks whose claims and key assumptions do not necessarily contradict each other. This compatibility assumption is too weak for various reasons that have been presented in this commentary so far. For it is the purpose of the theory synthesis sketched here to enrich and refine the notion of enculturation and the associated theses defended by CI. Furthermore, to the extent that PP directly speaks to complex cognitive phenomena and learning, it benefits from the effort of CI to do justice to the socio-culturally shaped context in which these phenomena can be developed. This is to say that CI and PP can be directly referred to each other in ways that I have started to illustrate in this section.

Finally, from this it follows that both frameworks are more than just compatible – they are *complementary*. Taken together, they provide us with complex and far-reaching conceptual tools for investigating complex cognitive phenomena that are shaped by the individual’s immersion in its cognitive niche. Thus, the complementarity of CI and PP leads to a new integrative framework that I dub enculturated predictive processing (EPP).

2.4 Defending enculturated predictive processing

At first glance, the EPP framework might appear to be unwarranted. For prediction error

⁴ Thanks to Jennifer M. Windt for raising this point.

⁵ Thanks to an anonymous reviewer for helpful suggestions on this issue.

minimization could be construed as being a purely internalistic, brain-bound affair that does not leave any room for the idea that cognitive processes are constituted both by neuronal and extracranial bodily components that are normatively constrained, socially scaffolded, and deeply anchored in a socio-culturally structured environment.

First, consider a position that takes for granted that cognitive processes can be coherently described in terms of prediction error minimization, but which denies that cognitive processes are co-constituted by neuronal and bodily sub-processes operating on socio-cultural resources. Such a position is defended by [Jakob Hohwy \(2013, p. 240\)](#) who argues that “[...] many cases of situated and extended cognition begin to make sense as merely cases of the brain attempting to optimize its sensory input so it, as positioned over against the world, can better minimize error.” In particular, according to his interpretation of the prediction error minimization framework, “[...] the mind remains secluded from the hidden causes of the world, even though we are ingenious in using culture and technology to allow us to bring these causes into sharper focus and thus facilitate how we infer to them.” (*ibid.*, p. 239)

For Hohwy, this directly follows from the causal relations holding between the predictive system and the environmental causes it constantly tries to infer. According to him (*ibid.*, p. 228), this relation needs to be characterized as “direct” and “indirect” at the same time:

[...] the intuition that perception is indirect is captured by its reliance on priors and generative models to infer the hidden states of the world, and the intuition that perception is direct is captured by the way perceptual inference queries and is subsequently guided by the sensory input causally impinging on it.

Since the causal relation that holds between a predictive system comprised of inverted generative models and the world is partly indirect, so the argument goes, the system is in constant embodied interaction and direct contact with its

environment only insofar as it tries to make the effects of hidden causes fit the predictions. This precludes the theoretical possibility of depicting prediction error minimizing systems as being situated, scaffolded, integrated, or extended.

However, this line of reasoning fails to acknowledge the conceptual necessity of emphasizing the functional role of embodied active inference in terms of its contribution to the minimization of prediction error and the optimization of predictions. For even if the causal relations holding between a predictive, generatively organized system and environmental causes are mediated by hypotheses, predictions, prediction errors and precision estimation as encoded in the cortical hierarchy, it does not follow that this system is just a passive receiver of sensory input that informs it about remote states in the environment. Similarly, it does not necessarily follow from the prediction error minimization framework that it “[...] creates a sensory blanket – the evidentiary boundary – that is permeable only in the sense that inferences can be made about the causes of sensory input hidden beyond the boundary”, as [Hohwy \(2014, p. 7\)](#) claims. Rather, the predictive system is part of its socio-culturally structured environment and has many possibilities for bodily acting in that environment in order to facilitate its own cognitive processing routines. Considering embodied active inference, it turns out that the causal relation holding between embodied action (in terms of bodily manipulation) and changes of the set of available stimuli in the environment is as direct as any causal relation could be. This is because these changes are an immediate effect of these very prediction error-minimizing and precision-optimizing actions, which in turn contribute to the performance of cognitive tasks. Furthermore, we need to take into account that genuinely human cognitive processes occur in a culturally sculpted cognitive niche, which is characterized by mathematical symbol systems, representational writing systems, artifacts, and the like, and other human organisms with whom we interact. These cognitive resources have unique properties that render them particularly useful for the completion of cognitive tasks.⁶ For example, consider the regularity of line

⁶ Thanks to Richard Menary for raising this important point in personal communication.

arrangements and the orderliness of succeeding letters in an alphabetic writing system. Once learned and automatized, following these normative principles facilitates several types of cognitive processing routines. That is to say that it is the socio-culturally shaped sensory input itself that has an important impact on the concrete realization of prediction error minimization. This cannot be accounted for if we assume that the predictive processing of cognitive resources is an internalistic, secluded endeavour.

Second, consider a line of reasoning that goes against the compatibility of CI with the prediction error minimization framework, that might be put forward by an integrationist. She might agree that we need a mechanistic description of the neuronal and bodily components which jointly constitute cognitive processes in the close interaction with socio-cultural resources. But she might continue to argue that the performance of cognitive practices is more than just the minimization of prediction error and the optimization of precision.⁷ From the perspective of PP, it needs neither to be denied that human cognitive systems as a whole aim to fulfil cognitive purposes by completing cognitive tasks and that they do so by engaging in cognitive practices. Nor should it be rejected that cognitive practices are normatively constrained and that cognitive systems are deeply immersed in a socio-culturally structured environment, which in turn provides these very norms through scaffolding teaching. However, the important theoretical contribution made by the prediction error minimization framework is its providing of a sub-personal, mechanistic description of the underlying neuronal and bodily sub-processes that turns out to be parsimonious, conceptually coherent, and empirically plausible. In addition, PP also offers a description of the close interaction of the neuronal and bodily components constituting cognitive practices by offering a concise description of the ongoing, mutually constraining interplay of perceptual and active inferences. More generally, this section should have established that all important claims and assumptions made by CI in favour of cognitive

practices, such as the hybridity, the transformative efficacy, and the enculturated nature of cognitive processes, can be supplemented and refined by taking the prediction error minimization framework into account.

The arguments in favour of the EPP framework directly speak to the current debate within philosophy of mind and philosophy of cognitive science about the relationship between the prediction error minimization framework and approaches to situated, distributed, integrated, or extended cognition. On the one hand, [Jakob Hohwy](#) (2013, 2014) denies on both methodological and metaphysical grounds that there is anything like these types of cognition from the perspective of prediction error minimization. According to him, this is because predictive systems have only indirect access to the world. Furthermore, there is “the sensory boundary between the brain and the world” which prohibits predictive systems from engaging in any variant of situated, distributed, integrated, or extended cognition including CI ([Hohwy 2013](#), p. 240). On the other hand, [Andy Clark](#) (2013, p. 195) argues that the PP framework at least “[...] offers a standing invitation to evolutionary, situated, embodied, and distributed approaches to help ‘fill in the explanatory gaps’ while delivering a schematic but fundamental account of the complex and complementary roles of perception, action, attention, and environmental structuring.” Once we take the arguments and considerations in favour of EPP into account we have reasons to think that EPP lends support to Clark’s construal of the PP framework. This will become even more persuasive once we take empirical data and a paradigm case of EPP into account.

3 Reading acquisition: A case of enculturation

So far, I have argued that the notion of enculturation and key claims made by CI can be enriched by taking the PP framework into account. In particular, the hybridity, embodiedness, and transformative character of enculturated cognition can be mechanistically described in terms of prediction error minimization. How-

⁷ This consideration was put forward by Richard Menary in personal communication.

ever, cognitive practices cannot be fully reduced to prediction error minimization, since they have a normative dimension that needs to be investigated on a personal level of description.

This section serves to illustrate the validity of the line of reasoning put forward in this commentary. This will be done by showing that reading acquisition, understood as another case of enculturation next to mathematical cognition, can be fruitfully described from the perspective of EPP.

3.1 Scaffolded learning and the acquisition of cognitive norms

One crucial aspect of learning to perform a cognitive practice is the acquisition of the relevant cognitive norms, where this class of norms “govern[s] manipulations of external representations, which aim at completing cognitive tasks” (Menary 2010, p. 238). In the case of reading, these norms concern the recognition and identification of tokens of a representational writing system. In alphabetic writing systems, important cognitive norms are derived from the so-called *alphabetic principle*, where this principle amounts to the “mapping [of] written units onto a small set of elements – the phonemes of a language” (Rayner et al. 2001, p. 33; see also Snowling 2000, p. 87). Specifically, the correspondence of graphemes to phonemes puts culturally established, normative constraints on the ways in which individual letters (and combinations thereof) are related to phonological units. The normative scope of these correspondences is best illustrated by differences across languages and orthographies. As pointed out by Ziegler & Goswami (2006, p. 430), “[i]n some orthographies, one letter or letter cluster can have multiple pronunciations (e.g. English, Danish), whereas in others it is always pronounced in the same way (e.g. Greek, Italian, Spanish).”⁸ This demonstrates that the degree of consistency or transparency of *grapheme-phoneme correspondences* is subject to arbitrary stipulations by a linguistic, literate community employing a specific orthographic system. These stipulations are

normative insofar as they constrain the ways in which combinations of letters are pronounced and written words are correctly related to spoken words. The acquisition of this normative knowledge needs “explicit instruction in the alphabetic principle” (Rayner et al. 2001, p. 57).⁹ It follows that learning these norms is socially structured and dependent upon the cooperation of experts with novices. This fits neatly with Menary’s (2013, p. 361) following assumption:

Manipulative norms and interpretative norms apply to inscriptions of a public representational system and are never simply dependent on an individual. Indeed, it is the individual who must come to be transformed by being part of the community of representational system users.

Acquiring knowledge about grapheme-phoneme correspondences, especially in an inconsistent orthography such as English, puts demands not only on the novice, but also on the teachers who assist her in learning these correspondences. For the teachers, being experts in reading, need to break down their automatic identification and recognition skills in order to be able to teach the norms underlying the relationship between graphemes and phonemes. As Sterelny (2012, p. 145) points out more generally, “[e]xpert performance is often rapid and fluent, without obvious components. Learning from such performance is difficult. It becomes much easier if the task is overtly decomposed into segments, each of which can be represented and practiced individually.” In the present context, the most successful strategy of teaching grapheme-phoneme correspondence has turned out to be so-called *phonics instruction* (cf. Rayner et al. 2001, pp. 31f): “[...] teaching methods that make the alphabetic principle explicit result in greater success among children trying to master the reading skills than methods that do not make it explicit” (ibid., p. 34). This goes along with teaching novices that spoken language consists of phonemes. That is, children’s reading acquisi-

⁸ This phenomenon is also known as orthographic depth. For a recent review, see Richlan (2014).

⁹ See also Dehaene (2010, p. 219), Dehaene (2011, p. 26), and Frith (1985, p. 307).

tion is dependent upon, or at least co-develops with *phonological awareness*, where this is understood as “[...] the ability to perceive and manipulate the sounds of spoken words” (Castles & Coltheart 2004, p. 78). The *metalinguistic awareness* that spoken language consists of phonemes must be explicitly acquired and allows the novice to learn that these units correspond to letters, or combinations thereof. It is still debated whether phonological awareness is a prerequisite for learning to read or whether it is co-emergent with basic letter decoding skills. However, as suggested by Castles & Coltheart (2004, p. 104), “[...] it may not be possible for phonemic awareness to be acquired at all in the absence of instruction on the links between phonemes and graphemes.” Thus, it seems safe to assume that phonological awareness clearly facilitates the ability to relate graphemes to phonemes. There are other components of metalinguistic awareness that influence the successful application of norms governing alphabetic representational writing systems. Beginning readers are already proficient speakers of their native language and are able to fluently apply syntactic, semantic, and pragmatic norms in their everyday conversations. However, they are usually unable to explicitly represent that utterances are made up of sentences and that sentences are made up of combinations of words (cf. Frith 1985, p. 308; Rayner et al. 2001, p. 35). To novices, these basic properties must be made explicitly available in order to put those novices in the position to apply knowledge about them automatically and fluently at later stages of reading acquisition. Furthermore, novices need to be acquainted with the convention, which is fairly obvious to expert readers, that alphabetic writing systems are decoded from left to right and from the top to the bottom of a page. These basic personal-level components of the acquisition of reading skills provide the cognitive norms necessary for the development of reading understood as a cognitive practice. It is these norms that govern the successful manipulation of representational vehicles belonging to an alphabetic writing system that need to be established by social interaction between learners and teachers. Thus, be-

coming proficient in applying the alphabetic principle, getting to grips with phoneme-grapheme correspondences, and developing phonological and metalinguistic awareness are cases of scaffolded learning.

3.2 Reading acquisition and neuronal transformation

Next to scaffolded learning, another crucial aspect of cognitive transformation is LDP (cf. Menary 2013, p. 356, [this collection](#), p. 8). Indeed, in the case of reading acquisition, there is unequivocal evidence pointing to “[...] plastic changes in brain function that result from the acquisition of skills” (Ansari 2012, p. 93). By the same token, Ben-Shachar et al. (2011, p. 2397) emphasize that “[...] culturally guided education couples with experience-dependent plasticity to shape both cortical processing and reading development.” As Schlaggar & McCandliss (2007, p. 477) point out, the application of knowledge about grapheme-phoneme correspondences in novice readers “[...] implicates the formation of functional connections between visual object processing systems and systems involved in processing spoken language.” The left ventral occipitotemporal (vOT) area appears to play a crucial role in establishing these connections.

As mentioned by Menary ([this collection](#)), there has been consensus on the contribution of the vOT area to a neuronal reading circuit. In a series of experiments, Stanislas Dehaene, Laurent Cohen and their colleagues have made the remarkable discovery that neuronal activation in one particular region of the left vOT area is reliably and significantly associated with visual word recognition in adult, non-pathological readers (Cohen & Dehaene 2004; Dehaene 2005, 2010; Dehaene & Cohen 2011; Dehaene et al. 2005; McCandliss et al. 2003; Vinckier et al. 2007). This region, especially the left ventral occipito-temporal sulcus next to the fusiform gyrus, frequently responds to visually presented words regardless of the size, case, and font in which they are made available (cf. Dehaene 2005, p. 143; McCandliss et al. 2003, p. 293). This consistent finding has led these researchers

to call it the visual word form area (VWFA), since it crucially contributes to “[...] a critical process that groups the letters of a word together into an integrated perceptual unit (i.e. a ‘visual word form’)” (McCandliss et al. 2003, p. 293). However, it is debatable whether the left vOT area is almost exclusively dedicated to visual word recognition in expert readers, or whether this area serves several functions having to do with the (visual) identification of shapes more broadly construed (see Price & Devlin 2003, 2004, for a discussion). Nevertheless, the findings by Dehaene and his colleagues that the left vOT area plays a crucial role in the overall visual word recognition process is important and widely acknowledged, although the interpretations of its functional contribution differ.

An important motivation for research on the overall function of the left vOT area stems from considerations on the phylogenetic development of visual word recognition. Considering that writing systems were invented only approximately 5400 years ago, it is unlikely that the ability to read is the result of an evolutionary process (cf. Dehaene 2005, p. 134, 2010, p. 5; McCandliss et al. 2003, p. 293). In a nutshell, the crucial question is how visual word recognition is possible given “[...] that the human brain cannot have evolved a dedicated mechanism for reading” (Dehaene & Cohen 2011, p. 254). This is also referred to as the “reading paradox” (Dehaene 2010, p. 4). The solution to this paradox proposed by Dehaene and his colleagues is to assume “[...] that plastic neuronal changes occur in the context of strong constraints imposed by the prior evolution of the cortex” as a result of the human organism being exposed to tokens of a certain writing system (Dehaene & Cohen 2011, p. 254). Specifically, the idea is “[...] that writing evolved as a recycling of the ventral visual cortex’s competence for extracting configurations of object contours” (*ibid.*). This view, which has been dubbed the *neuronal recycling hypothesis* (cf. Dehaene 2005, p. 150), suggests that existing neuronal functions associated with visual cognition are “recycled” for the phylogenetically recent, ontogenetically

acquired capacity to recognize visually presented words (cf. Cohen & Dehaene 2004, p. 468; see also Menary 2014, p. 286). This “recycling” is in turn constrained by the overall evolved neuronal architecture and already existing processing mechanisms (cf. Dehaene 2010, pp. 146f). Thus, neuronal recycling is just a special type of neuronal reuse (see Anderson 2010, for a discussion). There are certain conditions that need to be met if a specific cortical area is to be ‘recycled’ for a phylogenetically recent cognitive function (see Menary 2014, p. 288). In the case of visual word recognition, the left vOT area is assumed to exert certain “functional biases” that make it most suitable for the recognition and identification of visually presented words: “(1) a preference for high-resolution foveal shapes; (2) sensitivity to line configurations; and (3) a tight proximity, and, presumably, strong reciprocal interconnection to spoken language representations in the lateral temporal lobe” (Dehaene & Cohen 2011, 256). These “functional biases”, however, do not preclude that the left vOT area is still engaged in other cognitive processes such as object recognition in skilled adult readers (cf. Carreiras et al. 2014, p. 93; Dehaene & Cohen 2011, p. 257; Price & Devlin 2004, p. 478). Rather, it helps explain why this area is found to be well-equipped for contributing to the overall process of visual word recognition. However, the question arises what the contribution of the left vOT area to the overall visual word recognition process is supposed to make. According to Cathy Price’s & Joseph Devlin’s (2011) Interactive Account (IA), the contribution of the left vOT area can be best described and explained in terms of PP. In line with the general principles of the PP framework presented above, they generally hold the following assumption: “Within the hierarchy, the function of a region depends on its synthesis of bottom-up sensory inputs conveyed by forward connections and top-down predictions mediated by backward connections” (Price & Devlin 2011, p. 247). In other words, the suggested synthesis equals the prediction error that results from the discrepancy

between top-down predictions and bottom-up sensory information. Applied to the patterns of neuronal activation associated with visual word recognition, this assumption is specified as follows:

For reading, the sensory inputs are written words (or Braille in the tactile modality) and the predictions are based on prior association of visual or tactile inputs with phonology and semantics. In cognitive terms, vOT is therefore an interface between bottom-up sensory inputs and top-down predictions that call on non-visual stimulus attributes. (Price & Devlin 2011, p. 247)

Accordingly, the vOT area is supposed to be associated with a distinct level of the hierarchical generative model responsible for visual word recognition mediating between higher-level, language-related predictions and bottom-up visual information. It follows that “[...] the neural implementation of classical cognitive functions (e.g. orthography, semantics, phonology) is in distributed patterns of activity across hierarchical levels that are not fully dissociable from one another” (*ibid.*, p. 249). Specifically, IA proposes a neuronal mechanism that is able to demonstrate how linguistic knowledge about phonology and semantics, encoded in top-down predictions, causally interacts with bottom-up information. This is because it is held that a prediction error is generated each time bottom-up information diverges from the associated top-down prediction. In turn, the resulting prediction error is associated with significant activation in the left vOT area. Empirical evidence supporting this approach to the functional contribution of the left vOT area to visual word recognition in expert readers is widely available (see, e.g., Bedo et al. 2014; Kherif et al. 2011; Kronbichler et al. 2004; Schurz et al. 2014; Twomey et al. 2011).

In reading acquisition, the left vOT area appears to be an equally important contributor to visual word recognition. According to Price & Devlin (2011, p. 248), the activation level of the vOT area develops in a non-linear fashion,

as the proficiency in visual word recognition increases:

In pre-literates, vOT activation is low because orthographic inputs do not trigger appropriate representations in phonological or semantic areas and therefore there are no top-down influences [...]. In early stages of learning to read, vOT activation is high because top-down predictions are engaged imprecisely and it takes longer for the system to suppress prediction errors and identify the word [...]. In skilled readers, vOT activation declines because learning improves the predictions, which explain prediction error efficiently [...].

That is, IA assumes that the level of activation within the left vOT area is dependent upon the general establishment and refinement of a generative model comprising both lower-level areas associated with visual processing and higher-level cortical areas associated with phonological and semantic knowledge. If this account turns out to be correct, the blurredness of the distinction between perception and cognition as suggested by Clark (2013) becomes vitally important. For it is the mutual interplay of lower-level processing stages (traditionally associated with visual processing) and higher-level processing stages (traditionally associated with phonological and semantic processing) that renders the successful acquisition of visual word recognition possible in the first place. Evidence in favour of IA comes from studies demonstrating that there is a significant increase of activation in this area as a result of exposure to visually presented words in beginning readers across different research paradigms and methodologies employing fMRI (e.g., Ben-Shachar et al. 2011; Gaillard et al. 2003; Olulade et al. 2013). Furthermore, two longitudinal ERP studies (Brem et al. 2010; Maurer et al. 2006) demonstrate that the left-lateralized occipito-temporal N1 effect, an effect associated with print sensitivity, does not develop in a linear fashion in the course of reading acquisition. Rather, Maurer et al.’s (2006, p. 756) comparison of their results obtained from their child participants with an adult control

group indicates that “[i]nstead of a linear increase with more proficient reading, the development is strongly nonlinear: the N1 specialization peaks after learning to read in beginning readers and then decreases with further reading practice in adults following an inverted U-shaped developmental time-course.” In this vein, [Brem et al. \(2010, p. 7942\)](#) interpret their results by suggesting that “[t]he emergence of print sensitivity in cortical areas during the acquisition of grapheme-phoneme correspondences is in line with the inverse U-shaped developmental trajectory of print sensitivity of the ERP N1, which peaks in beginning readers [...]”

Another consequence of [Price’s & Devlin’s \(2011\)](#) PP account of reading acquisition is that the activation level within the vOT should be associated with the degree of accuracy of top-down predictions in the face of bottom-up signals. This is supported by various studies demonstrating that higher-level activations of cortical areas associated with language processing are also present in beginning readers. For example, [Turkeltaub et al. \(2003, p. 772\)](#) report that “[a]ctivity in the left ventral inferior frontal gyrus increased with reading ability and was related to both phonological awareness and phonological naming ability. [...] Brain activity in the anterior middle temporal gyrus also increased with reading ability”, where this area is associated with semantic processing. Similarly, [Gaillard et al. \(2003\)](#) report activation in the middle temporal gyrus, which is frequently associated with semantic processing in expert readers (e.g., [Bedo et al. 2014, p. 2](#); [Price & Mechelli 2005, p. 236](#); [Vogel et al. 2013, p. 231](#); [Vogel et al. 2014, p. 4](#)). Furthermore, they report significant activation patterns in left IFG, which is associated with both phonological and semantic processing.

In the light of much empirical evidence in favour of [Price’s & Devlin’s \(2011\)](#) approach to the neuronal changes corresponding to reading acquisition, it seems safe to assume that it is empirically plausible and can account for many data derived from experiments in cognitive neuroscience. However, to what extent can this approach be conceptually enriched? Recall that learning a new skill such as reading is just a

special case of overall prediction error minimization according to the PP framework. On this construal, learning to read means becoming increasingly efficient in predicting linguistic, visually presented input as a result of long-term exposure to types of this input and the optimization of hypotheses through perceptual inference. The careful instruction in relating graphemes to phonemes, phonological and metalinguistic awareness, and the normatively constrained alphabetic principle provides the environmental conditions for efficient and progressively more accurate prediction error minimization. The signals delivered by this highly structured learning environment are estimated as being precise, such that the synaptic gain on error units reporting the discrepancy between (still inaccurate) predictions and prediction error is high. As learning to read proceeds, the predictions become more accurate and the overall influence of prediction error shows a relative decrease. This line of reasoning is supported by [Price’s & Devlin’s \(2011, p. 248\)](#) following suggestion: “At the neural level, learning involves experience-dependent synaptic plasticity, which changes connection strengths and the efficiency of perceptual inference.” Understood this way, LDP and the associated neuronal transformations can be understood as being realized by prediction error minimization in the context of scaffolded learning, which allows a beginning reader to become ever more efficient and successful in this particular cognitive practice.

3.3 Reading acquisition and bodily transformation

Starting from the hybrid mind thesis defended by CI, which states that certain cognitive processes are constituted by both neuronal and extracranial bodily sub-processes, it seems natural to assume that reading acquisition also is associated with the transformation of bodily sub-processes. That is, in the course of enculturation it is the enactment of bodily manipulation that is transformed in addition to the neuronal changes occurring as a result of LDP. In terms of PP, this assumption leads to the suggestion that it is not only perceptual inferences that are

causally relevant for learning described in terms of prediction error minimization, but also active inferences that allow for ever more efficient sub-personally employed strategies for “explaining away” incoming sensory input. Recall that eye movements are just a special case of active inference (see e.g., [Friston et al. 2012](#)). Their functional contribution to prediction error minimization becomes vitally important for a complete account of visual word recognition and its acquisition. This is because visual word recognition, in both novices and experts, is rendered possible by the coordination of perceptual and active inference. From the perspective of CI, the idea here is that the ways in which an individual bodily manipulates a certain cognitive resource is importantly improved in the course of cognitive transformation. Applied to reading acquisition, this leads to the assumption that specific eye movement patterns become more efficient as a result of reading instruction and iterate exposure to a certain type of cognitive resource (say, sentences printed on a piece of paper).

Recently, it has become possible to investigate eye movements in beginning readers by employing eye-tracking methodologies. Converging evidence suggests that beginning readers make more fixations (i.e., acquisition of visual information in the absence of oculomotor activities), saccades (i.e., oculomotor activities), and regressions (i.e., backward saccades), and exhibit longer fixation durations and smaller saccade amplitudes than proficient and expert readers (cf. [Joseph et al. 2013](#), p. 3; [Rayner et al. 2001](#), p. 46). More specifically, these tendencies are assessed in a longitudinal eye-tracking study reported by [Huestegge et al. \(2009\)](#). They measured eye movements during an oral reading task in second and fourth graders of a German primary school and additionally assessed overall reading skills and oculomotor behaviour beyond reading (cf. [Huestegge et al. 2009](#), p. 2949). Their results indicate that the fourth graders, in comparison to the second graders, show a decrease of fixation duration, gaze duration, total reading time, refixations, and saccadic amplitudes (cf. [ibid.](#), p. 2956). [Huestegge et al. \(2009](#), p. 2958) attest that the younger, less

proficient readers show a “[...] refixation strategy, with initial saccade landing positions located closer to word beginnings.” Similarly to [Huestegge et al. \(2009\)](#), [Seassau et al. \(2013\)](#) report a longitudinal study comparing the performance of 6- to 11-year-old children in a reading task and a visual task. In line with the empirical evidence already mentioned, their results indicate that “[w]ith age, children’s reading capabilities improve and they learn to read by making larger progressive saccades, fewer regressive saccades and shorter fixations [...]” ([Seassau et al. 2013](#), p. 6). Furthermore, it is demonstrated that the eye movement patterns employed in reading and in visual search diverge with increasing reading proficiency (cf. [ibid.](#), p. 9).

An explanation of these results in terms of PP is straightforward. In beginning readers, the predictions initiating active inference occurring in a highly-structured linguistic environment are inaccurate, such that the generation and execution of eye movements in terms of active inference is not as efficient as it is in the case of expert readers. By the same token, the inaccuracy of the currently selected prediction makes it necessary to sample the visually available linguistic environment more thoroughly, explaining the “refixation strategy” and the execution of comparatively more saccades. As reading skills improve, resulting from increasingly efficient prediction error minimization through perceptual inference as already suggested, the accuracy of predictions becomes increasingly optimal, therefore allowing for more efficient active inference. More efficient active inference, in turn, allows for more efficient perceptual inference, since both types of inference mutually influence each other. This line of reasoning is supported by [Huestegge et al.’s \(2009](#), p. 2957) claim informed by the results of their study “[...] that only linguistic, not oculomotor skills were the driving force behind the acquisition of normal oral reading skills.” Thus, the increase in efficiency of eye movements in beginning readers does not result from an increase in oculomotor capabilities *per se*, but works in tandem with higher-level linguistic knowledge encoded in predictions, which are associated with representa-

tions in higher-order cortical areas. As a result, the improvement of active inference in the course of reading acquisition works in tandem with the improvement of perceptual inference. This highlights that learning to read does not only result in neuronal, but also in bodily transformations. As such, the optimization of eye movements in the course of reading acquisition highlights the importance of bodily manipulation in the efficient enactment of reading understood as a cognitive practice. This also means to suggest that a complete account of enculturation should not only pay attention to scaffolded learning and LDP, but also to the developmental trajectory of bodily manipulation.

4 Concluding remarks

This commentary on Richard Menary's paper *Mathematical Cognition: A Case of Enculturation* started from the assumption that the general outline of enculturation and the associated claims made by CI provide important conceptual tools for the description of ontogenetically acquired, socio-culturally shaped cognitive processing routines. However, I have argued that the idea of enculturation and its most important aspects, namely cognitive transformation and scaffolded learning, need to be enriched by providing a detailed functional and neuronal description on a sub-personal level of description. In addition, it needs to be born in mind that enculturation is rendered possible by normative constraints developed by a large group of individuals sharing the same cognitive niche. To this end, I have suggested that the notion of enculturation and its associated constitutive aspects can be complemented in important ways by taking the PP framework into account. The result is what I call enculturated predictive processing. Thus, the PP framework is capable of providing the conceptual resources necessary for a thorough description of the mechanistic underpinnings of cognitive practices and their acquisition. Lending further support to this line of reasoning, I have dealt with reading acquisition as a paradigmatic case of enculturated predictive processing. This should have been sufficient to establish that the CI framework is well-suited

for a conceptually coherent description of the interaction between brain, body, and environmental cognitive resources. However, it needs to be supplemented by a sub-personal level description in terms of prediction error minimization in order to be able to specify the neuronal and functional underpinnings of the hybrid mind thesis, the bodily manipulation thesis, and the transformation thesis as defended by CI. At the same time, the approach to reading acquisition put forward in this commentary suggests that a vast array of empirical findings from cognitive neuroscience and cognitive psychology can be unified for the first time by interpreting them from the new perspective of enculturated predictive processing. Thus, I submit that we can only appreciate the cognitive assets rendered possible by our socio-culturally structured environment once we account for the enabling conditions of sophisticated, neuronally and bodily realized cognitive processes such as mathematical cognition and reading. These conditions include socio-culturally established ways of learning and teaching, LDP, and the ability to adapt action patterns to the needs and requirements of a certain cognitive task. My overall claim is that we need the EPP framework to be able to approach the entire spectrum of these factors, whose complex interplay ultimately leads to truly enculturated cognition.

Acknowledgements

The author wishes to thank the Barbara Wengeler Foundation for its generous financial support. In addition, she is indebted to Thomas Metzinger, Jennifer M. Windt, and an anonymous reviewer for their helpful feedback on earlier versions of this commentary.

References

- Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences*, 33 (04), 245-266. [10.1017/S0140525X10000853](https://doi.org/10.1017/S0140525X10000853)
- Anderson, M. L. & Finlay, B. L. (2014). Allocating structure to function: The strong links between neuroplasticity and natural selection. *Frontiers in Human Neuroscience*, 7. [10.3389/fnhum.2013.00918](https://doi.org/10.3389/fnhum.2013.00918)
- Ansari, D. (2012). Culture and education: New frontiers in brain plasticity. *Trends in Cognitive Sciences*, 16 (2), 93-95. [10.1016/j.tics.2011.11.016](https://doi.org/10.1016/j.tics.2011.11.016)
- Anscombe, G. E. M. (1963). *Intention* (2nd ed.). Cambridge, Mass: Harvard University Press.
- Austin, J. L. (1953). How to talk: Some simple ways. *Proceedings of the Aristotelian Society* (53), 227-246.
- Bedo, N., Ribary, U., Ward, L. M. & Valdes-Sosa, P. A. (2014). Fast dynamics of cortical functional and effective connectivity during word reading. *PLoS ONE*, 9 (2), e88940-e88940. [10.1371/journal.pone.0088940](https://doi.org/10.1371/journal.pone.0088940)
- Ben-Shachar, M., Dougherty, R. F., Deutsch, G. K. & Wandell, B. A. (2011). The development of cortical sensitivity to visual word forms. *Journal of Cognitive Neuroscience*, 23 (9), 2387-2399. [10.1162/jocn.2011.21615](https://doi.org/10.1162/jocn.2011.21615)
- Brem, S., Bach, S., Kucian, K., Guttorm, T. K., Martin, E. & Lyytinen, H. (2010). Brain sensitivity to print emerges when children learn letter-speech sound correspondences. *Proceedings of the National Academy of Sciences*, 107 (17), 7939-7944. [10.1073/pnas.0904402107](https://doi.org/10.1073/pnas.0904402107)
- Carreiras, M., Armstrong, B. C., Perea, M. & Frost, R. (2014). The what, when, where, and how of visual word recognition. *Trends in Cognitive Sciences*, 18 (2), 90-98. [10.1016/j.tics.2013.11.005](https://doi.org/10.1016/j.tics.2013.11.005)
- Castles, A. & Coltheart, M. (2004). Is there a causal link from phonological awareness to success in learning to read? *Cognition*, 91 (1), 77-111. [10.1016/S0010-0277\(03\)00164-1](https://doi.org/10.1016/S0010-0277(03)00164-1)
- Clark, A. (2012). Dreaming the whole cat: Generative models, predictive processing, and the enactivist conception of perceptual experience. *Mind*, 121 (483), 753-771. [10.1093/mind/fzs106](https://doi.org/10.1093/mind/fzs106)
- (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36 (03), 181-204. [10.1017/S0140525X12000477](https://doi.org/10.1017/S0140525X12000477)
- (2015). Embodied prediction. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Cohen, L. & Dehaene, S. (2004). Specialization within the ventral stream: The case for the visual word form area. *NeuroImage*, 22 (1), 466-476. [10.1016/j.neuroimage.2003.12.049](https://doi.org/10.1016/j.neuroimage.2003.12.049)
- Dehaene, S. (2005). Evolution of human cortical circuits for reading and arithmetic: The “neuronal recycling” hypothesis. In S. Dehaene, J.-R. Duhamel, M. D. Hauser & G. Rizzolatti (Eds.) *From monkey brain to human brain. A Fyssen foundation symposium* (pp. 133-157). Cambridge, MA: MIT Press.
- (2010). *Reading in the brain: The new science of how we read*. New York, NY: Penguin Books.
- (2011). The massive impact of literacy on the brain and its consequences for education. *Human neuroplasticity and education. Pontifical Academy of Sciences*, 117, 19-32.
- Dehaene, S. & Cohen, L. (2011). The unique role of the visual word form area in reading. *Trends in Cognitive Sciences*, 15 (6), 254-262. [10.1016/j.tics.2011.04.003](https://doi.org/10.1016/j.tics.2011.04.003)
- Dehaene, S., Cohen, L., Sigman, M. & Vinckier, F. (2005). The neural code for written words: A proposal. *Trends in Cognitive Sciences*, 9 (7), 335-341. [10.1016/j.tics.2005.05.004](https://doi.org/10.1016/j.tics.2005.05.004)
- Feldman, H. & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, 4. [10.3389/fnhum.2010.00215](https://doi.org/10.3389/fnhum.2010.00215)
- Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360 (1456), 815-836. [10.1098/rstb.2005.1622](https://doi.org/10.1098/rstb.2005.1622)
- (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11 (2), 127-138. [10.1038/nrn2787](https://doi.org/10.1038/nrn2787)
- Friston, K., Adams, R. A., Perrinet, L. & Breakspear, M. (2012). Perceptions as hypotheses: Saccades as Experiments. *Frontiers in Psychology*, 3. [10.3389/fpsyg.2012.00151](https://doi.org/10.3389/fpsyg.2012.00151)
- Frith, U. (1985). Beneath the surface of developmental dyslexia. In K. E. Patterson, J. C. Marshall & M. Coltheart (Eds.) *Surface dyslexia. Neuropsychological and cognitive studies of phonological reading* (pp. 301-330). Hillsdale, NJ: Erlbaum.
- Gaillard, W., Balsamo, L., Ibrahim, Z., Sachs, B. & Xu, B. (2003). fMRI identifies regional specialization of neural networks for reading in young children. *Neurology*, 60 (1), 94-100. [10.1212/WNL.60.1.94](https://doi.org/10.1212/WNL.60.1.94)
- Hohwy, J. (2011). Phenomenal variability and introspective reliability. *Mind & Language*, 26 (3), 261-286. [10.1111/j.1468-0017.2011.01418.x](https://doi.org/10.1111/j.1468-0017.2011.01418.x)

- (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology*, 3. [10.3389/fpsyg.2012.00096](https://doi.org/10.3389/fpsyg.2012.00096)
- (2013). *The predictive mind*. Oxford, UK: Oxford University Press.
- (2014). The self-evidencing brain. *Noûs*. [10.1111/nous.12062](https://doi.org/10.1111/nous.12062)
- (2015). The neural organ explains the mind. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Huestegge, L., Radach, R., Corbic, D. & Huestegge, S. M. (2009). Oculomotor and linguistic determinants of reading development: A longitudinal study. *Vision Research*, 49 (24), 2948-2959. [10.1016/j.visres.2009.09.012](https://doi.org/10.1016/j.visres.2009.09.012)
- Joseph, H. S. S. L., Liversedge, S. P. & Paterson, K. (2013). Children's and adults' on-line processing of syntactically ambiguous sentences during reading. *PLoS ONE*, 8 (1), e54141-e54141. [10.1371/journal.pone.0054141](https://doi.org/10.1371/journal.pone.0054141)
- Kherif, F., Josse, G. & Price, C. J. (2011). Automatic top-down processing explains common left occipitotemporal responses to visual words and objects. *Cerebral Cortex*, 21 (1), 103-114. [10.1093/cercor/bhq063](https://doi.org/10.1093/cercor/bhq063)
- Kronbichler, M., Hutzler, F., Wimmer, H., Mair, A., Staffen, W. & Ladurner, G. (2004). The visual word form area and the frequency with which words are encountered: Evidence from a parametric fMRI study. *NeuroImage*, 21 (3), 946-953. [10.1016/j.neuroimage.2003.10.021](https://doi.org/10.1016/j.neuroimage.2003.10.021)
- Maurer, U., Brem, S., Kranz, F., Bucher, K., Benz, R., Halder, P., Steinhausen, H.-C. & Brandeis, D. (2006). Coarse neural tuning for print peaks when children learn to read. *NeuroImage*, 33 (2), 749-758. [10.1016/j.neuroimage.2006.06.025](https://doi.org/10.1016/j.neuroimage.2006.06.025)
- McCandliss, B. D., Cohen, L. & Dehaene, S. (2003). The visual word form area: Expertise for reading in the fusiform gyrus. *Trends in Cognitive Sciences*, 7 (7), 293-299. [10.1016/S1364-6613\(03\)00134-7](https://doi.org/10.1016/S1364-6613(03)00134-7)
- Menary, R. (2006). Attacking the bounds of cognition. *Philosophical Psychology*, 19 (3), 329-344. [10.1080/09515080600690557](https://doi.org/10.1080/09515080600690557)
- (2007a). *Cognitive integration: Mind and cognition unbounded*. New York, NY: Palgrave Macmillan.
- (2007b). Writing as thinking. *Language Sciences*, 29 (5), 621-632. [10.1016/j.langsci.2007.01.005](https://doi.org/10.1016/j.langsci.2007.01.005)
- (2010). Cognitive integration and the extended mind. In R. Menary (Ed.) *The extended mind* (pp. 227-243). Cambridge, MA: MIT Press.
- (2013). The enculturated hand. In Z. Radman (Ed.) *The hand, an organ of the mind. What the manual tells the mental* (pp. 561-593). Cambridge, MA: MIT Press.
- (2014). Neural plasticity, neuronal recycling and niche construction. *Mind & Language*, 29 (3), 286-303. [10.1111/mila.12051](https://doi.org/10.1111/mila.12051)
- (2015). Mathematical cognition: A case of enculturation. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Menary, R. & Kirchhoff, M. (2014). Cognitive transformations and extended expertise. *Educational Philosophy and Theory*, 1-14. [10.1080/00131857.2013.779209](https://doi.org/10.1080/00131857.2013.779209)
- Olulade, O. A., Flowers, D. L., Napoliello, E. M. & Eden, G. F. (2013). Developmental differences for word processing in the ventral stream. *Brain and Language*, 125 (2), 134-145. [10.1016/j.bandl.2012.04.003](https://doi.org/10.1016/j.bandl.2012.04.003)
- Price, C. J. & Devlin, J. T. (2003). The myth of the visual word form area. *NeuroImage*, 19 (3), 473-481. [10.1016/S1053-8119\(03\)00084-3](https://doi.org/10.1016/S1053-8119(03)00084-3)
- (2004). The pro and cons of labelling a left occipitotemporal region "the visual word form area". *NeuroImage*, 22 (1), 477-479.
- (2011). The interactive account of ventral occipitotemporal contributions to reading. *Trends in Cognitive Sciences*, 15 (6), 246-253. [10.1016/j.tics.2011.04.001](https://doi.org/10.1016/j.tics.2011.04.001)
- Price, C. J. & Mechelli, A. (2005). Reading and reading disturbance. *Current Opinion in Neurobiology*, 15 (2), 231-238. [10.1016/j.conb.2005.03.003](https://doi.org/10.1016/j.conb.2005.03.003)
- Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D. & Seidenberg, M. S. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest*, 2 (2), 31-74. [10.1111/1529-1006.00004](https://doi.org/10.1111/1529-1006.00004)
- Richlan, F. (2014). Functional neuroanatomy of developmental dyslexia: The role of orthographic depth. *Frontiers in Human Neuroscience*, 8. [10.3389/fnhum.2014.00347](https://doi.org/10.3389/fnhum.2014.00347)
- Roepstorff, A. (2013). Interactively human: Sharing time, constructing materiality. *Behavioral and Brain Sciences*, 36 (03), 224-225. [10.1017/S0140525X12002427](https://doi.org/10.1017/S0140525X12002427)
- Rowlands, M. (1999). *The body in mind: Understanding cognitive processes*, Cambridge studies in philosophy. Cambridge, UK: Cambridge University Press.
- Schlaggar, B. L. & McCandliss, B. D. (2007). Development of neural systems for reading. *Annual Review of Neuroscience*, 30 (1), 475-503. [10.1146/annurev.neuro.28.061604.135645](https://doi.org/10.1146/annurev.neuro.28.061604.135645)
- Schurz, M., Kronbichler, M., Crone, J., Richlan, F., Klackl, J. & Wimmer, H. (2014). Top-down and bottom-up influences on the left ventral occipito-temporal cortex during visual word recognition: An analysis of effective connectivity. *Human Brain Mapping*, 35 (4), 1668-1680. [10.1002/hbm.22281](https://doi.org/10.1002/hbm.22281)

- Seassau, M., Bucci, M.-P. & Paterson, K. (2013). Reading and visual search: A developmental study in normal children. *PLoS ONE*, 8 (7), e70261-e70261. [10.1371/journal.pone.0070261](https://doi.org/10.1371/journal.pone.0070261)
- Seth, A. K. (2015). The cybernetic Bayesian brain: From interoceptive inference to sensorimotor contingencies. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Snowling, M. J. (2000). *Dyslexia*. Malden, MA: Blackwell Publishers.
- Sterelny, K. (2012). *The evolved apprentice: How evolution made humans unique, The Jean Nicod lectures: Vol. 2012*. Cambridge, MA: The MIT Press.
- Turkeltaub, P. E., Gareau, L., Flowers, D. L., Zeffiro, T. A. & Eden, G. F. (2003). Development of neural mechanisms for reading. *Nature Neuroscience*, 6 (7), 767-773. [10.1038/nn1065](https://doi.org/10.1038/nn1065)
- Twomey, T., Kawabata Duncan, K. J., Price, C. J. & Devlin, J. T. (2011). Top-down modulation of ventral occipito-temporal responses during visual word recognition. *NeuroImage*, 55 (3), 1242-1251. [10.1016/j.neuroimage.2011.01.001](https://doi.org/10.1016/j.neuroimage.2011.01.001)
- Vinckier, F., Dehaene, S., Jobert, A., Dubus, J. P., Sigman, M. & Cohen, L. (2007). Hierarchical coding of letter strings in the ventral stream: Dissecting the inner organization of the visual word-form system. *Neuron*, 55 (1), 143-156. [10.1016/j.neuron.2007.05.031](https://doi.org/10.1016/j.neuron.2007.05.031)
- Vogel, A. C., Church, J. A., Power, J. D., Miezin, F. M., Petersen, S. E. & Schlaggar, B. L. (2013). Functional network architecture of reading-related regions across development. *Brain and Language*, 125 (2), 231-243. [10.1016/j.bandl.2012.12.016](https://doi.org/10.1016/j.bandl.2012.12.016)
- Vogel, A. C., Petersen, S. E. & Schlaggar, B. L. (2014). The VWFA: It's not just for words anymore. *Frontiers in Human Neuroscience*, 8. [10.3389/fnhum.2014.00088](https://doi.org/10.3389/fnhum.2014.00088)
- Ziegler, J. C. & Goswami, U. (2006). Becoming literate in different languages: similar problems, different solutions. *Developmental Science*, 9 (5), 429-436. [10.1111/j.1467-7687.2006.00509.x](https://doi.org/10.1111/j.1467-7687.2006.00509.x)

What? Now. Predictive Coding and Enculturation

A Reply to Regina E. Fabry

Richard Menary

Regina Fabry has proposed an intriguing marriage of enculturated cognition and predictive processing. I raise some questions for whether this marriage will work and warn against expecting too much from the predictive processing framework. Furthermore I argue that the predictive processes at a sub-personal level cannot be driving the innovations at a social level that lead to enculturated cognitive systems, like those explored in my target paper.

Keywords

Active inference | Cognitive integration | Enculturation | Learning driven plasticity | Mathematical cognition | Perceptual inference | Predictive processing | Reading

Author

[Richard Menary](#)

richard.menary@mq.edu.au
Macquarie University
Sydney, NSW, Australia

Commentator

[Regina E. Fabry](#)

fabry@students.uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Introduction: What? Now.

I'd like to thank Regina Fabry for her excellent and detailed response to my paper. She articulates an important account of reading acquisition as a process of enculturation and describes how a Cognitive Integration/Enculturated Cognition (henceforth CI/ENC) account can be combined with a predictive processing account of neural processing. She shows, in impressive detail, how CI/ENC can benefit from Predictive Processing (henceforth PP), primarily as a way of explaining the neural-level details of processes that conspire with bodily interactions with the local environ-

ment to complete cognitive tasks. Since Fabry's response suggests an important way of cashing out some of the details of an enculturated approach, I would like to take this opportunity to look at some of the potential pitfalls in the proposed Enculturated Predictive Processing style (henceforth EPP). Primarily I want to focus on the differences in explanatory emphasis between CI/ENC and PP, especially where CI/ENC proposes the importance of the population-level effects of normative patterned practices (henceforth NPP), such as mathematical practices.

PP is all about predictions, happening in the here-and-now¹; however CI/ENC occurs at different levels and over much longer time-scales. It turns out that this difference is important, because if the brain is engaged in predictive error minimization (as sub-personal processing) in the here-and-now, then it cannot be driving the innovation of new NPP over many generations. This is because the pressures driving those innovations are found at the social, or populational, level², not at the level of neural processing where ‘what?’ is answered in the now.

I also raise several issues concerning the nature of the PP project, particularly whether, as a theory of general brain architecture, all processing can be cashed out in terms of predictive processes. I’m also sceptical about Fabry’s claim that PP can provide the “mechanistic underpinnings of the acquisition of cognitive practices” (Fabry this volume, p. 3) on its own, without help from what I call learning-driven plasticity (LDP) and neural redeployment. Finally, I comment on the promising research path down which Fabry is headed.

In the first section I remind the reader of some of the leading ideas of the CI/ENC framework, highlighting, in particular, the different levels of explanation and how this matters to the proposed marriage of ENC-PP. In the second section I raise several problems for the PP approach in general and for the ENC-PP approach in particular. My main concerns are to push away from an ‘isolated brain’ interpretation of PP and to place EPP within a much broader context of explanation.

2 CI and enculturation

As I point out in my contribution to this volume, cognitive integration should be understood as a thesis about the enculturation of human cognition. It is a thesis about how phylogenetically earlier forms of cognition are built

upon by more recent cultural innovations (e. g., systems of symbolic representation). This results in a multi-layered system with heterogeneous components, dynamically interwoven into a co-operative of processes and states an integrated cognitive system (henceforth ICS). The co-ordination dynamics of the system are, at least in part, understood in terms of the physical dynamics of brain–body–niche interactions in real-time; however, they are also to be understood in terms of NPP that govern and determine those interactions (over time). NPP operate at both social/population levels and individual, even sub-personal, levels. They originate as patterns of activity spread out over a population of agents; consequently they should be understood primarily as public systems of activity and/or representation that are susceptible to innovative alteration, expansion, and even contraction over time. They are transmitted horizontally across generational groups and vertically from one generation to the³ next. At the individual level they are acquired, most often by learning and training, and they manifest themselves as changes in the ways in which individuals think, but also the ways that they act (intentionally) and the ways in which they interact with other members of their social group(s) and the local environment. NPP, therefore, operate at different levels (groups and individuals) and over different time-scales (intergenerationally and in the here-and-now).

Given this, it is clear that What? Now⁴ processes that reduce prediction errors on their own could not drive the innovation of NPP; nor could they determine the properties of NPP on their own. Less obviously, I would argue, they do not drive the acquisition of NPP, because scaffolded learning requires both a physically and temporally-structured learning environment and the capacity for functional changes to cortical circuitry to be driven by the structured learning environment. The mechanism of acquisition includes both neural and environmental processes working in concert and over long periods of ontogenetic time. What? Now processes may help us to understand the here-and-now

¹ I mean predictions on incoming sensory input relevant to immediate action in the environment.

² I think that these levels are real. There is a level of entire populations, social groups, individual organisms and there is a level of individual brains. Cognition takes place within and across (at least the final three) levels.

³ See my target article for examples.

⁴ Predictions on sensory input in the here-and-now.

processes by which we enact NPP; they may even tell us something about the neural mechanisms for learning and plasticity; but we should be wary of making prediction and error minimization the driving factors behind the why and how of enculturation.

Fabry's commentary focuses on the neural level, functioning in real-time, where the primary aim is to give a mechanistic account of how cognitive capacities can be transformed by learning and training in rich socio-cultural niches. Rather than looking at the origin of ICS in cultural inheritance, phenotypic plasticity, and learning driven plasticity, Fabry argues that a version of the PP framework can provide the neural mechanisms by which ICS are (partly) constructed. My contribution to this volume focused primarily on the origin of ICS in the recent cultural evolution of NPP and then explored how mathematical practices could be learnt and how this process of learning could drive functional changes to circuitry in the brain. Consequently, the CI/ENC framework pursues the phylogenetic and ontogenetic basis of the larger brain-body-niche nexus. What, though, of the neural mechanisms of transformation?

I don't agree with Fabry's starting premise that CI/ENC lacks a mechanism of transformation: the mechanism of transformation is learning-driven plasticity (LDP) with neural redeployment in a scaffolded learning environment. The fundamental plasticity of the brain explains the nature of neural transformations and why the brain is open to scaffolded learning driven by the environment. (E)PP does not have the resources to explain redeployment (this is a theme I take up in the next section). Why would it, since PP is not a framework for explaining redeployment. It might be the case that PP fits with a certain conception of scaffolded learning such as path-dependent learning, but I have yet to see a thorough working-through of the details and it's not clear to me that all scaffolded learning should be reduced to a predictive form of path-dependent learning.

Fabry claims that a dynamical systems approach to integration "does not spell out the mutual influence that neuronal and extra-cra-

nial bodily components have over each other" (2015, p. 3). The EPP approach is supposed to fill in the details here. However, I suspect that this judgement is made a little too quickly, because the dynamical systems description of brain-body-niche interactions is in one sense a higher-level description of those interactions. The dynamical interactions are described as being part of a larger system comprising brain, body, and niche. We can zoom in and focus upon the dynamics of brain or body, but we shouldn't confuse the dynamics of the brain for the dynamics of the overall system. I have highlighted and outlined the neural dynamics required for enculturation in a number of places. For example, in the account of body schema dynamics and in the case of NPP for symbolic cognition, I have outlined the case for dual component transformations (e. g., Menary 2007, pp. 78–83; 2010; 2013 and 2014). Let's take these two cases in order.

In a now famous series of studies, Maravita & Iriki (2004) studied the bimodal interparietal neurons in trained Japanese macaque brains. These neurons respond both to tactile stimulation on the hand (tactile receptive field) and visual stimuli in the same vicinity as the tactile receptive field (the visual receptive field). The visual receptive field was centred on the hand following it through space. When macaques were trained to use a rake to pull food towards them on a table, the observation that struck Maravita and Iriki was that when the macaques used the rake the receptive fields of the bimodal neurons extended along the axis of the rake, including its head. Iriki's interpretation of this is that "either the rake was being assimilated into the image of the hand or, alternatively, the image of the hand was extending to incorporate the tool" (Iriki & Sakura 2008, p. 2230). The extension of the body schema (receptive field) to include the tool happened only during active holding; it reduced to just the hand during inactivity. The interesting result of these experiments is that the existing body schema has the latent capacity to extend to incorporate the tool. LDP can be cashed out in terms of functional changes as the result of scaffolded learning even in the case of

macaques, let alone the notoriously plastic brains of humans.

Functional changes can be cashed out in terms of neural redeployment and cortical connectivity. Returning to the case of mathematical cognition, inherited systems for numerosity are evolutionary endowments; we can be reasonably sure of this because they are constant across individuals and cultures and they are shared with other species. The numerosity systems are “quick and dirty”; they are approximate and continuous, not discrete and digital. By contrast, discrete mathematical operations exhibit cultural and individual variation; there is a big difference between Roman numerals and Arabic numerals. They are subject to verbal instruction (they actually depend on language); one must learn to count, whereas one does not learn to subitise. Mathematics depends on cultural norms of reasoning (mathematical norms). The ability to perform exact mathematical calculations depends on the public system of representation and its governing norms. We learn the interpretative practices and manipulative practices as a part of a pattern of practices within a mathematics community, and these practices transform what we can do. They are constitutive of our exact calculative abilities. Mathematical practices get under our skins by transforming the way that our existing neural circuitry functions.

The relationship between the evolutionarily earlier system and the recent development of public mathematical systems, norms, and symbols comes down to the redeployment of the cortical territories that are dedicated to evolutionarily older functions by novel cultural artefacts (e. g., representations, tools). The transformation results in new connections between the frontal lobe for number-word recognition and association, the temporal lobe for the visual recognition of number form, and the parietal lobe for the approximate recognition of magnitudes across both left and right hemispheres (Dehaene 1997).

The deeply transformative power of our learning histories in the cognitive niche relates to the development of our capacities for understanding symbolic representations and for phys-

ically manipulating inscriptions in public space. In learning to understand symbols, the first transformation involves our sensorimotor abilities for creating and manipulating inscriptions (the transformation of the body schema). This is something we learn to do on the page and in the context of a learning environment, in public space, before we do it in our heads. Our capacities to think have been transformed, but in this instance they are capacities to manipulate inscriptions in public space.

It looks like PP can provide models of some of the fundamental processing principles at work at the sub-personal neural level, but it is not obvious that it would replace LDP and neural redeployment in the mechanism of transformation. However, Fabry may be right and PP may add another string to the bow of our understanding of how the brain exhibits the plasticity required for cognitive transformation. In that case it provides extra explanatory depth to the account of enculturation, but only as part of a much broader explanatory framework.

3 Some worries for enculturated predictive coding

Fabry provides a persuasive case for how PP could provide the neural underpinnings of enculturation. In this section, however, I will raise some problems for the proposed marriage of CI/ENC and PP. The main issues I will address are as follows:

1. The incompatibility of the isolated brain interpretation (Hohwy 2013) and the active inference interpretation (Clark 2013) of PP.
2. The attempt to explain all cognitive processing in terms of prediction error.
3. The redeployment of neural circuitry as not being explained by PP.
4. The role of NPP as not being explained by the reduction of prediction error.

1. Isolating the brain

If CI/ENC has one central commitment, it is that we should not think of cognition as isolated from the environment. And yet this is ex-

actly how we ought to understand the predictive brain, according to a prominent interpretation of the PP framework. Whenever the PP framework is introduced, it is almost always introduced in the following way: “Accounts of PP generally assume that human perception, action, and cognition are realized by *Bayesian probabilistic generative models* implemented in the human brain. Since *the human brain does not have immediate access to the environmental causes of sensory effects, it has to infer the most probable state of affairs in the environment* giving rise to sensory data” (Fabry 2015, p. 4; my emphasis). The two main motivations for the PP framework are that the brain is isolated from the environment and must make a best guess as to what it is perceiving, and that this kind of probabilistic inference-making results in internal (neurally realized) models of the environment. Putting aside the probabilistic nature of the inferences, this just is old-fashioned individualism. There is a perceptual interface to an environment of hidden variables; the internal system creates internal models (representations) of those hidden environmental variables, which then causally produce behaviour. The internal states must predict the external variables via sensory input, but they have no direct access to the causal ancestry of the sensory input. This form of individualism is used as an explanation for why models and predictions are required: “Because the brain is isolated behind the veil of sensory input, it is then advantageous for it to devise ways of optimizing the information channel from the world to the senses” (Hohwy 2013, p. 238). Hohwy describes the mind–world relation as “fragile” because of the isolation of the brain, and this is why active inference is required.

The saving grace of the PP framework, from the perspective of CI/ENC, is active inference. In Clark’s version of PP active inference and cultural props help to minimize prediction errors (Clark 2013); and because of this, there is a deep continuity between mind and world mediated by active inference and the cultural scaffolding of our local niche. Curiously, Hohwy agrees with Clark’s interpretation, but at a cost. Hohwy agrees that active inference and the cul-

tural scaffolding of the environment help to change sensory input so as to minimize prediction error, but also “by increasing the precision of the sensory input” (Hohwy 2013, p. 238). According to Hohwy, the primary role of PP is perceptual inference; as a matter of “second order statistics” active inference helps to optimise sensory input so that perceptual inference is less error-prone.

Note the cost. First, active inference and cultural scaffolding is relegated to the secondary role of reducing prediction error for the primary cognitive job of perceptual inference, which is carried out wholly by matching statistical models to sensory input in the brain. Second, Hohwy shows that this interpretation of active inference should be understood against the background of the isolated brain. “The key point I am aiming at here is that this is a picture that accentuates the indirect, skull-bound nature of the prediction error minimization mechanism” (Hohwy 2013, p. 238). Organizing and structuring our environments makes sense if the mind–world relation is fragile in the way that Hohwy presents it, and also because this structuring makes perceptual inference more reliable. I take it that Fabry and Clark would deny this interpretation of the role of active inference and cultural scaffolding. Indeed, Fabry denies Hohwy’s ‘isolationist’ interpretation in her commentary.

However, Fabry does so by playing up the roles of NPP, which go far beyond prediction minimization: “Furthermore, we need to take into account that genuinely human cognitive processes occur in a culturally sculpted cognitive niche. [...] These cognitive resources have unique properties that render them particularly useful for the completion of cognitive tasks” (Fabry 2015, p. 12). She also nods to the sub-personal, mechanistic role of PP in the entire brain–body–niche nexus: “[T]he important theoretical contribution made by the prediction error minimization framework is its providing of a sub-personal, mechanistic description of the underlying neuronal and bodily sub-processes” (Fabry 2015, p. 13). It is therefore not clear to me that PP does anything more than provide the functional details of *some* of the neural processing in the brain–body–niche nexus. It cer-

tainly should not be taken to provide a comprehensive account of what cognition is and why there is cultural scaffolding, or what its interesting cognitive properties are.⁵ It is to these issues that I shall now turn.

2. Everything is predicted

One of the main concerns with the PP approach is that it is used both to try to explain all of cognition and as an explanation of why there is cultural scaffolding. We've already seen a brief hint of this in Hohwy, Clark, and Fabry's work above.⁶ The first worry can be found in the expression of PP as originating in the free energy principle:

The free-energy considered here represents a bound on the surprise inherent in any exchange with the environment, under expectations encoded by its state or configuration. A system can minimise free energy by changing its configuration to change the way it samples the environment, or to change its expectations. These changes correspond to action and perception, respectively, and lead to an adaptive exchange with the environment that is characteristic of biological systems. This treatment implies that the system's state and structure encode an implicit and probabilistic model of the environment. (Friston & Stephan 2007, p. 417)

PP is primarily a model of the way in which top-down processing 'predicts' bottom up sensory input and which samples the environment to change its expectations. These correspond to perception and action respectively.⁷ However, it seems odd to build a cognitive theory on the basis of the prediction of sensory sig-

nals. This is because much of cognition is not about sensory signal prediction; nor about actions as sampling the environment. Indeed much of cognition isn't about 'prediction' at all. So whilst I agree that at least part of the mechanisms of cognition can be fruitfully modelled by PP, not all of them will be. In enculturated systems, the really important work is being done by the processing governed by normative patterned practices whose properties are understood primarily at the social or populational level. I agree that at the individual level, the mechanisms of ICS can partly be explained by PP, but the main explanatory work will not be a matter of predictions of sensory input⁸.

The examples from Landy & Goldstone (2007) may be partly explained by prediction errors, but again this only makes sense in the context of sensorimotor processing governed by mathematical norms. If the norms function as priors in the system, then this might help explain the errors made by the test subjects.

3. Phenotypic plasticity and neural redeployment

PP can't explain the redeployment of neural circuitry to new cognitive functions. And it is not supposed to, since this isn't the job it was designed to do. However, this is a considerable weakness if PP is supposed to be the primary mechanism of enculturation. I've already canvassed the reasons why in section 1.

4. NPP and prediction error minimization

Enculturated PP plays a role in the multi-layered and interwoven ICS, but it neither determines nor implements the entire system. My argument in this response has been that the dynamics of ICS are not determined by the predictive processing of parts of the system: if any-

⁵ CI/ENC provides just these motivations and details. Clark himself proposes that the PP framework "offers a standing invitation to evolutionary, situated, embodied, and distributed approaches to help 'fill in the explanatory gaps' while delivering a schematic but fundamental account of the complex and complementary roles of perception, action, attention, and environmental structuring" (Clark 2013, p. 195).

⁶ See also their contributions to [this volume](#).

⁷ There are also theories of attention based upon PP, but I won't address those here.

⁸ Thomas Metzinger has raised an interesting question for me here: whether there is continuity between the levels? My argument has been that there is continuity between the levels, but this continuity is made possible by NPP's, LDP and neural redeployment. PP explains how we make perceptual inferences about the environment and it might explain something about the hierarchical organisation of neural architecture. However, it should be seen as playing a role in the organisation and enculturation of the brain, not the *only* role.

thing PP is enslaved to the processing needs of the entire enculturated system. The PP framework takes perceptual inference as its primary mode of processing, which is the top-down matching of predictions to sensory input. However, it is not obvious that this is the right model for all cognitive processing, since it is not obvious that all cognitive processing is just a matter of predictions about sensory input, nor a hierarchically organised system which minimises prediction errors.

For example Hohwy (2013, p. 238) argues that “many of the ways we interact with the world in technical and cultural aspects can be characterized by attempts to make the link between the sensory input and the causes more precise (or less uncertain).” This would be a very impoverished account of the evolution of public systems of representation. Public systems of representation did not simply evolve to “make the link between the sensory input and the causes more precise (or less uncertain)”; this would be to ignore the social pressures that would have caused representational innovation.⁹ It might be true that the history of the refinement of notation has something to do with making input more easy to ‘predict’; however, this would not be an *ultimate* explanation for why there are notations in the first place, nor how they function in our cognitive lives. It *might* be a *proximal* explanation of the neural mechanisms for the processing of notations and as such, it might explain some of the causal conditions that explain how notations have developed, but it doesn’t explain the conditions under which notations evolved. For further reasons why see section 3.4 of my target article, on evolutionary novelty and uniqueness (this volume).

For example, the idea that the brain predicts the product of two numerals makes sense, and the surprise at a product too distant from the operands lends further credence. Remember

the example from section 4.1 of my target article (this volume) : $34 + 47 = 268$. However, it is not obvious that predictions will help with the second example: $34 \times 47 = 1598$. What is required in this instance is the serial working through of the multiplication according to an algorithm. Furthermore, this is not simply a case of sensory predictions: when it comes to recognising the numerals on the page in front of you, PP can explain top-down predictions about sensory input, but that is not at all the same thing as the working through of a mathematical problem. So mathematical cognition could not, it seems to me, be reducible to error minimization.

4 Conclusion: Where now?

Despite some of my concerns about how the PP framework can be interpreted and its relation to the CI/ENC framework, I think that Fabry’s account of the enculturation of reading using a hybrid of CI and EPP is really compelling. This leads me to think that an EPP account might be workable for other cases, such as mathematical cognition. Having said this, the division of labour between PP and evolutionary accounts of the origin of NPP and ICS must be in place. The role of scaffolded learning and neural re-deployment should not be replaced by error minimization processes. The ‘isolationist’ reading of PP should be resisted, and a more situated cognition friendly approach embraced. PP is a sub-personal account of neural processes that fits within a larger account of the brain–body–niche nexus. If one embraces CI/ENC then there’s more to the mind than What? Now.

Acknowledgements

Thanks to Thomas Metzinger and Jenny Windt for comments and to Regina Fabry for her excellent commentary.

⁹ I take it that Hohwy is claiming that cultural representations function so as to make perceptual inferences more precise. This would be another way of reducing socio-cultural phenomena to a role that is complementary to the brain, with the processing needs of the brain dictating the evolutionary path that culture must take. The externalist perspective takes it that there are social and cultural pressures that require cognitive innovations (sometimes even new phenotypes).

References

- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36 (3), 181-204.
[10.1017/S0140525X12000477](https://doi.org/10.1017/S0140525X12000477)
- Dehaene, S. (1997). *The number sense: How the mind creates mathematics*. London, UK: Penguin.
- Fabry, R. E. (2015). Enriching the Notion of Enculturation: Cognitive Integration, Predictive Processing, and the Case of Reading Acquisition - A Commentary on Richard Menary. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Friston, K. & Stephan, K. E. (2007). Free energy and the brain. *Synthese*, 159, 417-458.
[10.1007/s11229-007-9237-y](https://doi.org/10.1007/s11229-007-9237-y)
- Hohwy, J. (2013). *The predictive mind*. Oxford, UK: Oxford University Press.
- Iriki, A. & Sakura, O. (2008). The neuroscience of primate intellectual evolution: natural selection and passive intentional niche construction. *Philosophical Transactions of the Royal Society B*, 363, 2229-2241.
[10.1098/rstb.2008.2274](https://doi.org/10.1098/rstb.2008.2274)
- Landy, D. & Goldstone, R. L. (2007). How abstract is symbolic thought? *Journal of Experimental Psychology*, 33 (4), 720-733. [0.1037/0278-7393.33.4.720](https://doi.org/10.1037/0278-7393.33.4.720)
- Maravita, A. & Iriki, A. (2004). Tools for the body (schema). *Trends in Cognitive Sciences*, 8, 79-86.
[10.1016/j.tics.2003.12.008](https://doi.org/10.1016/j.tics.2003.12.008)
- Menary, R. (2007). *Cognitive integration: Mind and cognition unbounded*. London, UK: Palgrave Macmillan.
- (2010). Dimensions of mind. *Phenomenology and the Cognitive Sciences*, 9, 561-578.
[10.1007/s11097-010-9186-7](https://doi.org/10.1007/s11097-010-9186-7)
- (2013). The enculturated hand. In Z. Radman (Ed.) *The hand, an organ of the mind. What the manual tells the mental* (pp. 593-561). Cambridge, MA: MIT Press.
- (2014). Neuronal recycling, neural plasticity and niche construction. *Mind and Language*, 29 (3), 286-303. [10.1111/mila.12051](https://doi.org/10.1111/mila.12051)
- (2015). Mathematical Cognition - A Case of Enculturation. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.

Understanding Others

The Person Model Theory

Albert Newen

According to Interaction Theory (IT), neither Theory Theory (TT) nor Simulation Theory (ST) give an adequate account of how we understand others. Their shared defect, it is claimed, is that both focus on third-person observation of the other, and neglect the role of social interaction. While interaction theory is made to account for the latter, it has problems doing justice to explicit attributions of propositional attitudes, especially from an observational stance. The latter received a new explanation by the Narrative Practice Hypothesis (NPH) which focuses on story-based explanations and tends to underestimate the relevance of nonlinguistic intuitive understanding. In this paper, I first try to do justice to what is plausible about each of the four approaches by accepting that each account introduces one plausible epistemic strategy for understanding others, which leads us to a multiplicity view about the epistemic strategies for understanding others. But it will then be argued that an adequate theory of understanding others needs further adjustment and correction because we need to account for the fact that we usually understand others on the basis of specific background knowledge that becomes more enriched during our life; I thus propose Person Model Theory (PMT) as a fruitful alternative. On my account, understanding turns on developing “person models” of ourselves, of other individuals, and of groups. These person models are the basis on which we register and evaluate persons as having mental as well as physical properties. I argue that person models can be either implicitly represented or explicitly available. This is accounted for by describing two kinds of person model, corresponding to the two ways of understanding others; very early in life we develop implicit *person schemata*, where a person schema is an implicitly-represented unity of sensory-motor abilities and basic mental phenomena related to one human being (or a group of humans); and we also develop *person images*, where a person image is a unity of explicitly-registered mental and physical phenomena related to one human being (or a group). I argue that the person model theory has more explanatory power than the other candidates.

Keywords

Person image | Person model theory | Person models | Person schema | Simulation theory | Theory theory

1 Introduction

A key question for social cognition is: Can we provide an adequate theoretical analysis of the process of understanding other human beings? For over twenty years, there have been only two possible answers to this question—that offered by “Theory Theory”, and that of “Simulation Theory”. The central claim of TT is that one’s understanding of another essentially relies on a folk-psychological *theory*, where some take the position that the relevant folk psychology is in-

born (e.g., [Baron-Cohen 1995](#)), while others claim that it is acquired ([Gopnik 1993](#)). In contrast, ST holds that we understand others by means of *simulation* (e.g., [Goldman 2006](#)), where simulation can take place at two levels, referred to as low-level and high-level simulation ([Goldman 2006](#)). In recent years, however, it has become clear that both positions have significant limitations. One central problem is claimed to be that both TT and ST take a

Author

[Albert Newen](#)

albert.newen@rub.de

Ruhr-Universität Bochum
Bochum, Germany

Commentator

[Lisa Quadt](#)

lisquadt@students.uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

primarily observational stance towards the other when analysing understanding:¹ critics maintain that this observational stance is a nonstandard, intellectual perspective, and that in fact we are normally involved in *interaction* when we try to understand others. Developing this line of thought, [Gallagher's](#) interaction theory (2001) combines involvement in interaction with a direct perception thesis, such that we can directly perceive the mental states of others and do not have to infer them. Another alternative proposal is [Hutto's](#) *narrative* account of social understanding (2008), on which understanding others relies centrally on telling or understanding stories. These idealized positions are the bases for a wide range of mixed positions, with which I will engage shortly. Yet even if we consider only these idealized positions, a new central defect quickly becomes clear: namely, that these positions offer answers to rather different questions. Thus, in a first step, I aim to reorganize the field of the main positions and use this framework to situate my own view, which I refer to as the *person model theory* ([Newen & Schlicht 2009](#); [Newen & Vogeley 2011](#)): this account is characterized by the claim that we understand others by essentially relying on person models of individuals, or of groups.

2 Reshaping the field of positions by distinguishing central questions

The question “How do we understand other human beings?” has to be divided into several subquestions, the first of which is: What epistemic strategy do we adopt to register or assess the other's cognitive states? To reach any kind of assessment of the other we need to obtain information within a concrete situation. The second question is: Once obtained, how is this prior information stored and organized? This

second aspect is important, because we always rely on prior background knowledge in our assessments of others. One main defect of the debate thus far has turns on the failure to distinguish these two questions. The debate between the two classic positions, ST and TT, can roughly be described as a misunderstanding stemming from their dealing with different questions: while ST insists that the use of simulation is the standard epistemic strategy, TT insists that the prior information we have about others is organized as a folk-psychological theory. Concerning their main claims, these accounts are not in opposition. The opposition only becomes visible if for each account we consider their favoured answer to both questions. The classic opposition between ST and TT can then be described as follows: TT claims that the epistemic strategy relies upon theory-based inferences, and that the prior information is organized as a folk-psychological theory; while ST claims that the strategy for information-processing involves simulation (to put oneself into the other person's shoes) which draws only on my own experience as the source of data for simulation, leaving it open as to whether these data form a theory.

Before turning to the question of which information-processing strategy we use to understand others, I first provide a brief survey of the field. Thus, in addition to TT and ST, we have [Gallagher's](#) IT, which focuses only on the strategy question; it claims that we understand others through social interaction and/or by direct perception, i.e., we can directly perceive mental phenomena; we also have [Hutto's](#) account, which is given in terms of story-telling. Their more elaborate joint account combines these claims ([Gallagher & Hutto 2008](#)), maintaining that we can distinguish three epistemic strategies for understanding others, depending on the stage of cognitive development in ontogeny: direct perception in very early childhood, followed by interactional understanding, and finally narrative understanding ([Hutto 2008](#)). In contrast, my aim will be to show that we actually use a multiplicity of information-processing strategies to understand others, depending on the context; the proposed account, then, is even

¹ This is a simplified view. A closer look into [Gopnik & Meltzoff \(1997\)](#) shows that their version of TT accounts for interaction as part of the development of an understanding of action and agency (Chap. 5). But interaction is not accounted for in the further dimensions of understanding others. From a bird's eye view this characterization is not inadequate, although it needs qualification. As the reader will see, my person model theory integrates this initial understanding of action and agency as elements of forming implicit person models that at the beginning may not be rich and abstract enough to warrant being called a theory (see n. 6 below).

richer than the three strategies proposed by the joint account of Hutto and Gallagher.

3 The epistemic strategy for understanding others

3.1 What about simulation?

According to Goldman's (2006) elaborate simulation account, we must distinguish between low-level and high-level mindreading. "Mindreading", in his view, comprises all cases of evaluating the mental state(s) of another person that normally lead to a language-based attribution of a mental state to a person. In the case of high-level mindreading, this is

[...] mindreading with one or more of the following features: (a) it targets mental states of a relatively complex nature, such as propositional attitudes; (b) some components of the mindreading process are subject to voluntary control; and (c) the process has some degree of accessibility to consciousness. (Goldman 2006, p. 147)

The paradigmatic case of high-level mindreading is understanding another person's decision. Third-person attribution of a decision consists of:

- imagining propositional attitudes in a form of *enactment imagination*;
- using (the same) decision-making mechanisms (as in the first-person case);
- projecting the result of using that mechanism onto a third person by attributing a decision.

We can easily present cases in which these proposed essential steps are not involved. For (i), to understand a person suffering from a delusion of persecution, we are not able to deploy enactment imagination: Their case is just too different from our own experience. And the same may be true in cases of deep cultural difference. For (ii), if I have experience with the other person such that I know that he has idiosyncratic, non-rational decision-making habits

when making weekend plans, I can use this knowledge to model his decision and not my own decision-making apparatus, since I have experience that my own apparatus differs from his (at least concerning weekend plans). For (iii), grant for the sake of argument that we have a plausible candidate for the beliefs and desires of the other and we use this for enactment imagination as well as input for my own decision-making apparatus, thus reaching a decision to do action A. Then, according to Goldman, I should project this decision onto the other person. Yet there remains an essential gap, which is noted by Goldman but not adequately addressed by him: He observes the necessity of "quarantining" my idiosyncratic background beliefs if I want to come to an adequate projection of the decision to do action A. Suppose I am warranted in presupposing that the other wants an ice-cream, has money, and that there is a nearby cafeteria where he can get one: then the decision-making apparatus may come to the decision to buy an ice-cream. If, however, I am a person who is extremely parsimonious with money, then my own background desire to save money may prevent me from buying the ice-cream in the same situation, and so this intervenes and I do not attribute the decision to buy an ice-cream to the other. But it seems that the desire to save money is—often, at least—an idiosyncratic desire that I should not use in my projection. Yet how do I know which of my own beliefs and desires are idiosyncratic and do not relate to the person I aim to understand? To solve this problem, I must already possess some view about the attitudes of the other as compared to me; yet this was what we were aiming to understand. In general, then, Goldman's theory of high-level mindreading has difficulties even getting off the ground: It starts by making presuppositions about the beliefs and desires of the other person, where this is exactly what we were aiming to understand. The same problem appears again in the projection phase, as just illustrated. Thus, high-level mindreading is a very special case of simulating a decision of the other, specifically when I already know a lot about the other, which I can use as input. This leaves open the question of how we get this in-

formation at all. Goldman tries to account for problems of this kind by accepting the importance of inference-based strategies and the organization of the prior information in form of a theory. Thus he is no longer developing a pure simulation theory but rather a hybrid account. Nevertheless, the counterexamples are not rare but in fact quite typical, and thus they cast doubt on the typicality and pervasiveness of high-level simulation in mindreading decisions.

Goldman may, however, appeal to his strategy of low-level mindreading, which is characterized as an activity that is “comparatively simple, primitive, automatic, and largely below the level of consciousness” (2006, p. 113). Goldman uses as a paradigmatic case face-based recognition of emotion, and he makes an additional appeal to “mirror neurons”, proposing that mirror neurons are not only relevant in the case of understanding motor activities (in both observing and doing them) but also for recognizing mental phenomena like pain and disgust. The most elaborate case relevant to this area of discussion concerns the study of disgust: It has been shown that experiencing disgust and observing disgust are dependent on certain mirror neurons that are activated in both cases (Wicker et al. 2003). Yet what exactly can we learn from this observation? I develop a critical position on the explanatory potential of mirror neurons in two steps. First, I argue that if mirror neurons could provide us with the whole story of how we understand others, this story would not be given as a case of simulation. Second, I cite evidence that mirror neurons do not provide the core part of the story of understanding others in cases of understanding emotions. Let us start with criticism of the claim that low-level mindreading is a case of simulation. Here I mainly rely on lines of criticism worked out by Gallagher (2007), who claims that “simulation is a personal-level concept that cannot be legitimately applied to subpersonal processes” (p. 363). Even if we do not accept Gallagher’s claim, the two core features of simulation would be lacking in the case of resonance processes implemented by mirror neurons: There is neither a first-person perspective involved nor

a type of pretence that includes a projection from a first-person perspective to a third-person perspective: “Thus, according to ST, simulation involves the instrumental use of a first-person model to form a third-person ‘as if’ or a ‘pretend’ mental state. For subpersonal processes, however, both of these characterizations fail” (Gallagher 2007, p. 360). Why are mirror neurons not an essential part of understanding others? They represent a type of action or emotion that is independent from a first- or third-person perspective; but the distinction between self and other is an essential part of understanding others. Thus a simulation process cannot be fully captured in its essential aspects by the mirror-neuron processes (see Vogeley & Newen 2002).

This criticism of high-level and low-level mindreading does not imply that simulation processes never take place: rather, it suggests that it is only so-called high-level simulation that we can characterize as simulation, and also that it is implausible that simulation is the standard strategy for everyday understanding of others. The latter claim is also based on the observation that we often rely on automatic, intuitive understanding of others without any conscious considerations.

3.2 What about theory-based inferences?

The same general line of criticism can be developed with respect to theory-based inferences. Such inferences may sometimes be relevant, but are not always so; neither are they the standard strategy for understanding others. Theory-based inferences are important when we are confronted with cases that we find strange or surprising, i.e., situations where we meet another person suffering from a mental disease which we know nothing about, or where the person belongs to a culture that is radically different from ours. In such scenarios, we consciously build hypotheses about the relevant mental phenomena, as well as about the best behavioural strategy to adopt. But most everyday scenarios in which we understand others are not of this type; quite the contrary, we are generally in-

volved in well-known situations with individuals or types of persons with whom we are familiar. There is an effortless application of our know-how regarding dealing with other humans, without any need to rationalize through theory-based inferences. The reply of the advocate of TT would be: Even if the relevant knowledge-how does not involve an explicit theory-based inference, it is only applicable because we rely on *implicit theory-based inferences*. The criticism of this line of thought is twofold: The status of *implicit inferences* is very unclear, because inferences are defined as relations between propositions; and there is evidence that implicit information processes are often non-propositional in nature. For example, in the case of experts, very often the epistemic strategy in their field is complex visual pattern-matching without any inferences; with their superior organization of knowledge, for instance, a chess expert can rapidly perceive a promising move, or a medical expert can quickly notice an inconsistency in a suggested diagnosis. The process of smoothly using this information mainly relies on fine-grained pattern-discrimination and pattern-matching (Gobet 1997) in the relevant situation, rather than on drawing inferences (which only becomes the case if the expert has to consider problematic situations). This is supported by observations of the way people recall chess positions: When seeing a chess board that contains a real, meaningful arrangement, chess experts excel as compared to novices in recalling positions, but perform no better for scrambled, impossible positions (Gobet & Simon 1996). This indicates that they are able to “see” meaningful patterns that a novice cannot see. They may use this ability in addition to making inferences, but inferences are not so much their basic access strategy as an additional one.² If neither the strategy of simulation nor the strategy of theory-based inferences is the standard strategy upon which our smooth, everyday understand-

ing of others is based, what form does epistemic access to others’ mental states take?

3.3 What about direct perception?

In recent years Gallagher (2008) has argued that our epistemic access to others’ mental phenomena is essentially based on direct perception. The mental states of others are not hidden, and need not be inferred on the basis of perceiving others’ behaviour; rather, behaviour is an expression of the mental phenomena that, in seeing the behaviour, is also seen directly. What does the claim of direct perception involve? Gallagher explains his main idea with an analogy: I can directly see my car. It would be inadequate to claim that I only directly see the colour, the shape, and the material, and then have to infer that it is my car. This is also supported by the fact that, when seeing the car, I at the same time see its drivability. This view does not deny that object-perception involves complex and partially hierarchically-organized brain processes, but it introduces the notion of “smart” perception: If I have learned the concept CAR and I am used to driving cars, I can see a car directly; and in seeing my car I may also see concomitant affordances such as its drivability. The same is true in the case of understanding others: according to Gallagher, by seeing their face and body posture in a specific situation, I can directly see that someone fears an aggressive dog. This can be realized by visual pattern-matching without inferences (see footnote 3 and Newen et al. forthcoming). This is a convincing comparison, especially as regards its potential to give a unified account of both basic perception and what Gallagher calls “smart” perception. The latter are cases in which it appears plausible to accept that perception can be modulated by conceptual information, these usually being described as cases of cognitive penetration (see Macpherson 2012; Vetter & Newen 2014).

Let us illustrate both the basic and the smart perception of an emotion. Basic perception of an emotion takes place when we see fear, joy, anger, or sadness in the face of a person while relying mainly on a single feature, or

² It is important to note that I leave it open whether we have to rely on a package of knowledge we are warranted in calling a theory, since I only discuss the strategy of information processing, not the organization of prior knowledge in experts.

small group of features, connected with facial expression (Ekman et al. 1972).³ This can be done through a bottom-up perceptual process that involves almost no top-down influences, especially if the facial expression is very characteristic of an emotion pattern. In the case of smart perception, the perception of the emotion is modulated by higher-order cognitive processes. To show this, we need a case in which the same facial input leads to a different perception of an emotion as a result of conceptual input. Such cases have indeed been discovered: If we first hear a story describing a very unjust situation that makes us expect the person we are going to see to be angry, we have a strong tendency to see a typical “Ekman” fearful face as an angry face: for example, if I am told that the relevant person made a reservation at the restaurant, waited for an hour while many other people who had come in later were served first, and that after a further hour was informed that she would have to wait for at least another hour, then I have a strong expectation of seeing anger. This has been shown to make us see a typical fearful face as an angry face (Carroll & Russell 1996). Smart perception of an emotion is a cognitively-penetrated perception of an emotion, and it is also important for seeing more complex emotions that do not have the typical Ekman facial expressions: if I know that John is jealous of Peter, because he told me so, and I have seen several episodes of Peter behaving intimately towards John’s wife Anne, and the next day I see another episode of John flirting with Anne while Peter observes them, I can directly see the jealousy in Peter’s face. There is no need for inference-based evaluations. This is parallel to Gallagher’s case of seeing one’s car:

3 Although our basic perception mainly relies on certain central cues—e.g., wide-open eyes for fear—the fearful face is not recognized only in one central feature of the face. It requires the integration of several facial features, and not static ones alone. The perceiver also benefits from noticing dynamic visual features like gaze direction: If the gaze is directed away from the perceiver instead of towards her, then this makes the recognition of fear occur faster (see Adams & Kleck 2003; Sander et al. 2006). Together with colleagues I have argued elsewhere that emotion recognition is essentially a process of pattern recognition (Newen et al. forthcoming). This is true for these basic perceptions of emotions. The face is integrated with body posture, since facial expressions are categorized as expressing a specific emotion most rapidly when they are paired with emotionally congruent body postures (Meeren et al. 2005; van den Stock et al. 2007).

we may describe both cases as cases of *seeing as*: seeing my car as a car (by knowing which affordances come with it) and seeing John’s face as evincing jealousy. I illustrated these cases of direct perception because I think Gallagher makes an important point when he claims that the main source of understanding others is direct perception (whether basic or smart). Nevertheless, there are clear limits to direct perception as a form of epistemic access.

Although Gallagher has in the past shown a tendency to overgeneralize the importance of the role of direct perception (2008), he is well aware that there remain cases that cannot be accounted for without going beyond direct perception. This is the case especially concerning our understanding of propositional attitudes—e.g., someone’s desire to take a summer holiday with his elder brother in western Turkey. Propositional attitudes are normally radically underdetermined by expressive elements such as facial expressions, gestures, body postures, etc., in a given situation. In general, therefore, complex human cognitive phenomena of this underdetermined type are communicated by linguistic exchange, or else have to be inferred or simply guessed on the basis of available information. The latter often happens in situations of non-transparent communication due to norms in social situations, or due to the fact that at least one person wants to hide her beliefs and intentions. Since these situations are also part of our everyday life, inferential processes remain part of our everyday understanding of others. Thus, although direct perception is a very important epistemic strategy that we may use in cases of face-based perception of emotion, even “smart” direct perception is not the basic strategy employed to understand complex beliefs, desires, and intentions of others. The latter require inferential processes as well. Thus, we are left with three strategies (simulation, theory-based inferences, direct perception), where none is a clearly dominant standard strategy relevant to all mental phenomena.

But there is at least one further candidate we should take into account, namely *understanding through primary interaction* (Gallagher & Hutto 2008). All the epistemic strategies dis-

cussed so far can apply to situations in which I am simply observing the other without being involved in any interaction. As we have already mentioned, Gallagher views this as a radical defect of such accounts; intuitive understanding of others is part of our everyday life, and this is especially the case if I am not in a purely observational situation but am directly involved in some kind of interaction. Intuitive understanding may then be characterized just by the fact that I notice a social act being directed towards me and so start to interact, such that a standard social interaction is realized, which may be non-linguistic but may also involve linguistic communication—e.g., friendly greetings exchanged while arranging ourselves in line at the office coffee machine. Such a strategy of understanding can only be dominant if the interaction is situated within many conventions, such that smooth understanding can take place without theoretical considerations about the others' beliefs and intentions (de Bruin et al. 2012). But is understanding through primary interaction, as it already takes place in neonate imitation (Meltzoff & Moore 1977, 1994), really the main or the standard strategy for understanding others? Again, even if we grant that this is an important strategy in basic understanding of others, even in adults—e.g., in minimal understanding deployed by smoothly interacting with a stranger who is taking the same bus—we need more advanced strategies to frame estimations about the ramifications of the situation—e.g., whether taking this bus in an unknown city, by night, and with such people on board, is a reasonable risk to take.

3.4 The multiplicity view

To summarize thus far. We use at least four epistemic strategies to understand others, and we learn to use these strategies on the basis of evidence of successful application in the past in relevantly similar situations. We prefer to use simulation strategies where we have evidence that the other is similar to us in respect of many features that are relevant to the situation of evaluation. We typically use theory-based inferences if we need to account for complex men-

tal phenomena or if an intuitive understanding is, for whatever reason, not available. We use understanding by primary interaction in cases in which we are involved in interaction with the other and only need to understand her or him to a limited degree, such that acting according to conventions is sufficient for a smooth interaction. Finally, we normally rely on direct perception of mental phenomena when we are in an observational stance towards the other and have a rich, well-organized body of experience that allows us to recognize mental phenomena as patterns. This is rather easy in cases of emotion recognition, more complex in recognizing intentions, and almost impossible in understanding complex propositional attitudes of others. Only the combination of all four strategies, in full sensitivity to the context and applied on the basis of our experience in successfully using the strategies, makes us experts in understanding others. Thus, we have reached a first main conclusion concerning strategies of understanding, this being what I call the multiplicity view:

The multiplicity view =_{Def} There is no standard default strategy of understanding others, but in everyday cases of understanding others we rely on a multiplicity of strategies that we vary depending on the context and on our prior experiences (and which are eventually also triggered by explicit training).⁴

This thesis is also supported by a closer look at mental disorder such as Asperger's syndrome, which is a variant of autism (Fiebich & Coltheart under review). People with Asperger's syndrome lack an intuitive understanding of others. They are unable to directly perceive emotions on the basis of facial expressions, and they tend to avoid social interaction (Vogeley 2012). Thus intuitive understanding by primary interaction or direct perception is not available for them. Since they also tend to experience themselves as being different (Vogeley 2012), they do not use simulation as a strategy: so

⁴ This view was worked out in parallel by Anika Fiebich in her PhD thesis, under my supervision. She applied the thesis in discussing the case of autism (defended January 2013).

they are left principally with theory-based inferences (Kuzmanovic et al. 2011). And this is what we can observe: persons who are autistic try to understand others by asking for theoretical guidance; thus they might ask how long one is allowed to look into the eyes of another person (Kai Vogeley, personal communication; his expertise is based on regular treatment of more than 300 patients). They also learn what people think in typical situations, but become lost in new situations. Since we have to deal with new situations almost every day, autistic people notice their tendency to get lost and many of them avoid social encounters. This special situation is explained by the fact that in contrast to the usual multiplicity of strategies of understanding, they are left with theory-based inferences alone. People with Down's syndrome are in a contrary kind of situation: they have a good intuitive understanding of others' emotions, but, due to typically very constrained cognitive abilities, they lack any theory-based inferences. In the early years of childhood—where cognitive skills are not so important as in kindergarten or school—their social life is very similar to the social life of children without Down's syndrome; but in later life the interdependence of social interaction with cognitive abilities leads to more problems in building an inclusive social life (Buckley et al. 2002). Thus, the normal multiplicity of strategies may be strongly constrained in some conditions of mental disorders. Furthermore, we can roughly cluster direct perception and interaction as the main epistemic access for *an intuitive understanding* of others, while *inference-based understanding* is based mainly either on a (high-level) simulation strategy or theory-based inferences (including inferences from narratives, see below). Since in our everyday life most of what is going on is intuitive understanding of others, it is especially important to highlight the relevance of social perception. In what follows, I will argue that the most important unit of clustering information about others is neither a facial unit nor an emotion type (or some other sub-personal unit), but the *whole person*—and thus a primary aspect of epistemic access is our ability to perceive persons. We *perceive persons* and

their mental settings mainly by directly perceiving them, and/or interacting with them. In addition, we can come to *judgments regarding persons* by simulating them and/or through inference-based understanding.

4 The organization of relevant background knowledge about others

We can now address the second independent question concerning understanding others: How do we organize the information about other people that we already have? This question presupposes that in standard cases of understanding others we are not in a situation in which we are bereft of relevant background knowledge. Quite the contrary: most of the time, we interact with people about whom we have a lot of background knowledge—family members, colleagues, friends, etc. Furthermore, we have background knowledge about the general needs of human beings, the special needs of students, homeless people, etc. It seems clear that we are relying on this type of knowledge in an essential way when we understand others. There may be very short period as a newborn baby when we start from scratch, armed only with certain inborn minimal mechanisms such as neonate imitation. Even the social smile developed with two months is dependent on external stimulation and learning processes, and babies very quickly start to react selectively towards familiar and foreign individuals. They also expect a typical behavioural interactive pattern from the caregiver. If a mother stops reacting intuitively through normal facial expressions and gestures, and instead reacts with a “still face”, then the baby quickly starts to cry (Bertin & Striano 2006; Nagy 2008). The baby is irritated by the unexpected pattern of reaction. How, then, are all these different types of background information about the other organized and used in social understanding?

4.1 Are we organizing our prior knowledge in folk-psychological theories?

The question of whether we are organizing our knowledge according to folk-psychological theor-

ies has received a number of different answers. According to TT, this is exactly what happens. In understanding others we rely on folk-psychological rules such as: “If she desires an ice-cream and she believes that she can get one with her money at the cafeteria, then she will go to the cafeteria”. No doubt folk-psychological rules, organized according to a belief–desire psychology, are an important instrument for understanding others; but they are by no means the only one. Often it is sufficient to know the conventions in a society to understand what someone is doing and will do next, e.g., if someone is in Japan and he enters a restaurant, he will first take off his shoes, then take a seat, and then will be asked to order. So, seeing someone entering a restaurant who looks like a guest (and not a waiter) allows us to expect a specific conventionally-regulated sequence of behaviour. If one has a liberal notion of folk-psychological theory, then we may add such behavioural conventions into that theory. But even then the question remains whether our understanding of others always relies on knowledge organized as a folk-psychological theory. A counterexample can be proposed by reference to cases of basic intuitive understanding: e.g., the still-face reaction by the caregiver, instead of a typical smiling facial expression and gestural response, makes the baby start to cry (as we saw above). There is thus an intuitive recognition of basic emotions like fear, anger, happiness, or sadness. This may rely on inborn emotion recognition mechanisms, or mechanisms learned very early, which may be evolutionarily anchored, since recognizing such basic emotions is essential for survival (Griffiths 1997; Panksepp 2005). There are two ways in which the counterexample might be blocked: (i) It could be maintained that some folk-psychological theories are inborn (Baron-Cohen 1995) and that intuitive understanding such as face-based recognition of emotion already involves a theoretical package. The problem with this line of reasoning is that the notion of theory, stretched that far, starts to look very implausible. A theory is constituted by a minimal package of systematically interconnected beliefs; and even if a belief is understood in a liberal way such that it does not presuppose linguistic rep-

resentations, it remains highly questionable whether basic cases of face-based recognition can be characterized as a systematically interconnected set of beliefs. The standard descriptions of face-based recognition of emotion (e.g., Goldman 2006) on a neural level highlight the relevance of mirror neuron mechanisms and characterize the underlying mechanism as a rather basic and partially independent pattern-recognition process, and thus as not forming a theory. A defect in recognizing disgust does not automatically lead to a defect in recognizing other basic emotions like happiness or sadness (Calder et al. 2000). (ii) A more promising move is to claim that the folk-psychological theory is learned (Gopnik 1993). This view is compatible with some basic processes of understanding which do not yet form a theory, but are developed into one as they are integrated step by step into a systematically-organized body of knowledge. This is a plausible and to some extent empirically grounded view (Gopnik & Meltzoff 1997; Newen & Vogeley 2003).⁵ One shortcoming of this view, however, is that its proponents tend to appeal to examples that have a strong focus on general folk-psychological rules, such as: “All humans need to drink, thus if someone picks up a glass in the kitchen, he intends to pour into it some liquid to drink”. This neglects a very important phenomenon, namely that we mostly interact not with complete strangers but with persons we know at least partly and often very well. For example, if Michael observes his son in the kitchen grasping a glass he does not appeal to the folk-psychological rule at all, since he knows that his son—despite his education—still only drinks from a bottle when at home, and that if he takes up a

5 Gopnik and Meltzoff insist that the basic registration of objects—e.g., their being sensitive to object permanence, as well as the basic registration of agents rooted in their being able to distinguish inanimate objects and living beings—which babies develop very early on, shows that they already have an *initial theory* of objects and agents. They argue that the already innate “structures are rich enough and abstract enough to merit the name of theories themselves” (Gopnik & Meltzoff 1997, p. 82). But it is questionable whether the notion of theory really has any fruitful role here, because, for example, explanations and predictions of the behaviour of a baby when seeing an object are extremely constrained. The developmental story told by Gopnik and Meltzoff is of course very plausible and at some point turns into a theory, because the transformation of the representation in the context of new cognitive abilities comes with a rich and systematic package of explanations and predictions.

glass it is just because he wants to use it for practising magic tricks. This indicates that all the theories canvassed thus far have a blind spot: so far it seems simply to have been neglected that we rely extensively on knowledge of properties of individuals, which is organized as belonging to one specific individual (the son, the partner etc.) or to a group (students, managers, etc.). The general worry concerning the organization of this knowledge, according to TT, can also be expressed as follows: How are we able to apply a general theory of typically human features in a *specific social* situation? If we want to integrate our prior background knowledge of persons as individuals or as belonging to a group, e.g., to a profession, then we can characterize the organization of this knowledge as *person models*. Person models of individuals and groups are by far the most important source of understanding others, I will argue, and since they involve specific knowledge, they are the natural candidate for enabling adequate deployment of more general knowledge of human psychology in concrete everyday situations. It remains to be discussed, then, whether person models have the status of a folk-psychological theory or not. To adumbrate my line of argument: no doubt some elaborate person models are systematically-interconnected sets of beliefs, but not all of them have to be, because some person models only involve very sparse and basic properties that are not highly interconnected.

4.2 Do we organize our prior knowledge in narratives?

As we saw earlier, one recent account of understanding others, proposed by [Dan Hutto \(2008\)](#), holds that understanding others mainly relies on telling stories and using this knowledge to understand individuals. The core claim of his NPH (Narrative Practice Hypothesis) is

[...] that direct encounters with stories about persons who act for reasons—those supplied in interactive contexts by responsive caregivers—is the normal route through which children become familiar

with both (1) the basic structure of folk psychology and (2) the norm-governed possibilities for wielding it in practice, thus learning both how and when to use it. ([Hutto 2008](#), preface, p. x)

One focus of his theory is not so much how the prior background knowledge of others is organized, but rather how children are able to acquire it. His developmental claim is that the central route for learning relevant background knowledge is listening to stories about persons. I grant that this is an important additional route of epistemic access to relevant knowledge about others; but it is already an advanced method, not normally used before the second year of life. Furthermore, in such cases the focus is not epistemic access to knowledge used to understand the other in the situation (i.e., when listening to the storyteller), but rather to gain new background knowledge with an eye to future understanding of others. In a follow-up paper written together with Gallagher ([Gallagher & Hutto 2008](#)), Hutto and Gallagher enrich their views about epistemic access through appeal to direct perception and interaction (see above) in addition to learning by narratives. It is important to note the difference between epistemic access to information that allows me to understand the other in the actual situation (see section 3) and epistemic access to background knowledge relevant for future usage. Thus, by granting that narratives are an additional instrument for learning about important properties of persons, I can enrich my multiplicity claim as characterized above. In integrating this idea, one should also generalize it: we not only learn important background information that helps us to understand others by listening to stories told by a caregiver, but also by reading stories, especially novels.

Let us now briefly discuss the NPH considered as a claim about the organization of our background knowledge. If I have elaborate and explicit knowledge of a person, I may have acquired it by listening to or reading a story, and I may tell a story if someone asks me about this person. But, as the interaction view highlights, sometimes my knowledge may be anchored in

the interaction, yet still be non-linguistically represented, and only activated in similar interactive situations. Our rich non-linguistic knowledge about other human beings, which we acquire when directly perceiving them (tone of voice, what they look like) or interacting with them, or when realizing a joint action, etc., are often not linguistically coded and thus not memorized as a linguistic story. If we widen the notion of a story such that it includes any sequence of memorized events, we lose track of any interesting notion of “story”. In fact, we are instead going in the direction that I propose, i.e., that we organize our prior knowledge about others through unifying it in person models. Some such models may include properties of a person that are connected as or with stories, but the core of a person model is a unity of features of a person that are grouped together as belonging to one individual or to a group, where the features may be as primitive as the tone of voice of a person, and have no connection to any story, even in a wide sense.

Although our prior knowledge about others is the main component of our understanding of others in a specific situation, most of the theories canvassed above did not present any clear view on how this knowledge is organized.⁶ We found only two suggestions: relevant prior knowledge is organized either as a folk-psychological theory or as a narrative. Neither proposal covers all relevant cases: neither accounts for the innate or very-early-learned (nontheoretical) basic background knowledge that enables us to effect smooth interaction and allows us to rely on a basic intuitive understanding of others. And, furthermore, as I argue in the following, there is an alternative view, the person model theory, which is able to integrate the plausible aspects of these two suggestions, and additionally allows us to explain a variety of phenomena that the alternative views did not or cannot take into account—especially the integration of features of

other human beings that allow us to realize an intuitive understanding of them.

5 The person model theory

Before expounding the new account, let me highlight two main criteria of adequacy for any plausible candidate theory and some open questions. (i) The theory should account for two levels of understanding others from a phenomenological perspective, namely intuitive understanding and inference-based understanding. This was first clearly discussed by [Gallagher \(2001\)](#), while [Goldman \(2006\)](#) described it in his distinction between low-level and high-level mindreading. What, we may then ask, would be an adequate way of establishing this distinction? (ii) We learned from [Gallagher \(2005\)](#) that we should distinguish understanding others by observation from understanding by interaction.

There are also a number of open research questions that can potentially be answered in developing the alternative account: (a) What is the relation between understanding oneself and understanding others? Here the ST claims that understanding oneself is the basis for all understanding of others, while TT is neutral; [Carruthers](#), for example, has famously argued that understanding others is the source of our self-understanding ([2009](#)). (b) What is the relation between understanding persons and understanding objects or situations? (c) How can we best account for the difference between understanding a well-known person, on the one hand, and a complete stranger, on the other?

The new alternative theory, which promises to deal with these open questions, is the person model theory. The central claim of this theory is that we organize our prior knowledge that is used to understand others into something we can call person models, and that accounting for our way of using person models is the most informative factor when analyzing our everyday understanding of others. A person model⁷ is a unity of properties or features that

⁶ This includes, e.g., the ST, which mainly offers a claim about how we use our knowledge to understand others, and that the main source of this knowledge—in addition to situational input—is one’s own experience. But a representative of ST can easily grant that relevant prior knowledge is organized in a folk-psychological theory. She only insists that the strategy of application of this knowledge in a situation is a simulation process.

⁷ An important question which I cannot discuss in this paper is the question of the development of person model and the limits of application. Some very sketchy remarks may be of help here for urgent

Table 1: Varieties of person models

Person models	Self	Other: Individuals	Other: Groups
Person schema	Self schema	Individual person schema	Group person schema
Person image	Self image	Individual person image	Group person image

we represent in memory as belonging to one person or a group (resp. type) of persons. To account for the difference between two types of understanding others (intuitive versus inference-based understanding), I suggest that there are two types of person models in use: implicit person models,⁸ which we shall call person schemata; and explicit person models, which we shall call person images. Very early in life we develop *person schemata*: a person schema is an implicit person model and can typically be described as a unity of sensory-motor abilities and basic mental phenomena⁹ realized by basic representations and associated with one human being (or a group of humans), where the schema typically functions without any explicit considerations and is activated when directly seeing or interacting with another person. A person

schema is thus the unity of implicitly-available information about a person that is thus not easily accessible in terms of being reportable but is nevertheless used in a specific situation. In other words, a person schema is the basic unit that enables a practical knowledge (a *knowledge how*) for dealing with another human being while this ability relies mainly upon social perception and interaction. Person schemata can be developed step by step into *person images*. A person image is a unity of explicitly represented and typically consciously available mental and physical phenomena related to a human being (or a group of people). Thus, a person image is the unity of rather easily and explicitly available information about a person, including the person's mental setting. Both person schemata and person images can be developed for an individual, e.g., one's mother, brother, best friend, etc., as well as for groups of people, e.g., medical doctors, homeless people, managers, etc. Furthermore, person models are created for other people but also for oneself.¹⁰ In the case of modelling oneself we can speak of a self-model that we develop implicitly as a self-schema and explicitly as a self-image. Thus, we have the following varieties of person models (see Table 1).

Person models are characterized here as memorized units of person features, ignoring the difference between long-term or short-term memorization.¹¹ Person models are distinguished

questions: Concerning the development I suggest that person model unfolds gradually from an early model of living agents which is based on sensitivity for clusters of features indicating animacy and agency. This "agent models" enfold into person models which are systematically enriched by the features I describe as belonging to person schemata and person images. Furthermore, a creation of a person model (which is a unity of information clustered together) does not presuppose a concept of a person. Person models are developed in fact if some typical features of adult healthy human beings are clustered to model an individual or a group of entities which are relevantly similar to adult healthy human beings. Typical core features are e.g., 1. being an agent, 2. being a sentient being, 3. having some minimal control of action. We use person models to understand babies and pets since we usually perceive them as having a minimal amount of core features.

⁸ I am only presupposing a minimal consensus on using the distinction of implicit versus explicit. It indicates a (gradual) difference in epistemic access such that paradigmatic cases of explicit contents are easily accessible (by the subject's experience, memory, thinking, imagining etc.) while paradigmatic cases of implicit contents are very difficult to access by the subject while they nevertheless influence the subject's cognition and behaviour. Intuitively, explicit content are correlated with our intuitive understanding of *conscious accessibility*, but since the latter is scientifically pretty unclear, I do not want to ground the implicit/explicit distinction on the difference between being or not being consciously accessible.

⁹ Mental phenomena have different ontological types: states, events, processes, and dispositions. So not only are stable mental phenomena included but so are situational experiences (like tokens of perceptions, emotions, attitudes, etc.).

¹⁰ The distinction between *person schema* and *person image* is based on Shaun Gallagher's distinction between *body schema* and *body image*. Establishing a *person schema* of my own body amounts to Gallagher's *body schema*, while a *person image* of my own body is similar to what he introduces as *body image* (2005, p. 24).

¹¹ In a more detailed explication of the theory, it would indeed be useful to distinguish short-term person models (only stored in working memory) and long-term person models (stored in a long-term memory). In addition, other established distinctions in memory can be used to characterize the content of person models, such as procedural and declarative contents as well as episodic and semantic contents. I will, however, ignore these distinctions in this paper.

from the result of understanding in a situation, which may be either a person impression that mainly relies on person schemata, or a person judgment that mainly relies on person images. Let me illustrate one clear virtue of adopting the distinction between person schema and person image by reference to the fact that it can account for the difference between intuitive understanding and inference-based understanding of others.

5.1 Person schemata

In detail, then, what are person schemata? A *person schema* is an intuitively formed, implicit model of a person; it is a memorized unity of characteristic features of a person including facial features and expression, voice, moving pattern, body posture, gestures, and other perceivable features of a person. The function of clustering these features is to allow us to evaluate a person very quickly in a situation according to evolutionarily-important aspects: is a person familiar, dangerous, aggressive, helpful, or attractive? The evaluation is either expressed in a type of interaction, or it can simply be memorized in an implicit unitary structure for future retrieval, including recognizing the person and activating the former evaluation (Reddy 2008). Our main access to others in everyday life is through perceiving a person and forming an impression (see the review published as a book chapter by Macrae & Quadflieg 2010). To form a person impression, (i) we typically pick up these basic features by means of a quick visual evaluation, even when seeing a person for the first time, where (ii) most features are directly associated with socially-relevant information, and (iii) they are clustered at the level of perceiving the whole person. Let me offer some support for all three characteristics of the process of forming a person impression in a situation that is memorized as a person schema:

(i) *Quick evaluation even with parsimonious information*: Evaluations of threat (which is of strong evolutionary relevance) can be made on the basis of exposure to an unfamiliar face lasting as little as 39 milliseconds (Bar et al. 2006). If the exposure to the unfamiliar face

lasts about 100 milliseconds, we are able to evaluate likeability, trustworthiness, competence, and aggressiveness with subjective reliability levels that are similar to those generated under longer viewing times (Willis & Todorov 2006).¹²

(ii) *Most features are associated with socially relevant information*: looking into the face is a very rich source of information about a person. Between 3 and 7 months of age, infants learn to recognize the face of the mother and to distinguish it from the faces of strangers, and they start to categorize people according to emotional expression and sex (Nelson 2001). One important source of information that children use from 4 months onwards is the gaze-direction of a person, it having been shown that they can distinguish a direct from an averted gaze (Vecera & Johnson 1995). From 9 months onwards, infants learn to register the joint attention of the infant and an adult as directed towards an object (Cleveland & Striano 2007). Thus, on the basis of gaze-interaction they evaluate whether joint attention towards an object has been established or not, and learn to direct the attention of the other if necessary (Tommasello 1999). Between the ages of 9 and 18 months, children start to use gaze-information to register the *goal* of the action of the other human: they attend immediately to the eyes when the intentions of an actor are ambiguous (Phillips et al. 1992).

Let me now pick out some results based on studies of adults that illustrate the informational value of single cues. To start with facial expression: in emotion recognition, highly in-

¹² The time course can be observed in ERP studies. These studies all support claims about the early information processing of faces, although there is an ongoing debate about how best to interpret the results. The main observations are enhanced responsiveness to faces relative to a variety of other objects with peaks at approximately 100 milliseconds (Herrmann et al. 2005; Liu et al. 2002; Pegna et al. 2004), 170 milliseconds (Bentin et al. 1996; Eimer & McCarthy 1999; Itier & Taylor 2004), and 250 milliseconds (Bentin & Deouell 2000; Schweinberger et al. 2004) after stimulus onset. (For review see Macrae & Quadflieg 2010). Whole bodies (without faces) are evaluated with a delay of 20 milliseconds compared to the evaluation of faces (Gliga & Dehaene-Lambertz 2005). Concerning faces with emotional expressions, the following rather stable result is reported: there is a frontocentral positivity as early as 120 milliseconds after stimulus onset and a later more broadly distributed positivity beyond 250 milliseconds; both are modulated by emotional facial expressions (Eimer & Holmes 2002; Holmes et al. 2003; Vuilleumier & Pourtois 2007; Williams et al. 2006).

formative features include knitted eyebrows for sadness, a smile for happiness, and a frown for anger (Ekman 1972, 1999). To prevent this remark giving the wrong impression, I here highlight some individual features and will argue in the next step that they are part of an integrated view at the level of persons. Salient biological visual markers allow us to easily identify the “big three” categories in person perception (Brewer 1988; Fiske & Neuberg 1990), i.e., sex, race, and age. In the same way, we can illustrate highly informative single features such as body posture: if the other is bending her head in a communicative context, this is unconsciously registered as signalling sympathy (Frey 1999).¹³ One important data source here is biological motion-detection as investigated by point light studies. If a person has lights on her hands, feet, and ankles, and some other significant parts of her body, we can videotape her bodily movement in the dark. Such artificial pure biological movement information allows us to register social features, e.g., we can recognize emotions (Ambady & Rosenthal 1992) and attribute personality features (Heberlein et al. 2004) on the basis of seeing dynamic movements alone. Furthermore, there is evidence that social information can be taken from the combination of gesture and body posture alone. In an intercultural study (Bente et al. 2010), an interaction between an employer and an employee (played by two students of one type of culture) was filmed for a short period. Then the film was edited to show only gesture and body posture. This was realized by showing idealized wooden puppets, representing the real interaction while abstracting from facial information, speech, clothing etc. The question to be addressed was, what we can read from seeing the body postures and gestures. The interactions were filmed with students from UAE (United Arab Emirates), Germany, and the United States; and the test subjects were also drawn from all three countries. With this film, people could determine whether the people in the scene were nervous or not, as well as the dominance relation, i.e., they

could see who was the boss. This is an interculturally-shared social understanding of otherwise culturally variable cues of body posture and gesture (the US students moved a lot while the UAE students moved rarely). They furthermore could perceive the level of friendliness in the interaction, although the study showed that we are good at this only in assessing our own culture.¹⁴ Furthermore, there are many more complex culturally-dependent visual features that (according to other studies) we use for evaluating the other—e.g., physical attractiveness, where attractive people are evaluated as possessing more desirable characteristics than their less attractive counterparts, a phenomenon that has been labelled the *beauty-is-good stereotype* (Dion et al. 1972; Eagly et al. 1991). These kinds of stereotypes are especially connected with racial classifications: African-Americans are stereotypically assumed to be lazy, criminal, and uneducated, but also musical and athletic (Devine & Elliot 1995), whereas Asian-Americans are considered to be intelligent, industrious, conservative, and shy (Lin et al. 2005). Most observers in our culture assume that people with stylish hair and extravagant clothing are highly extrovert (Borkenau & Liebler 1992). We live with a lot of these deeply culturally-anchored stereotypes, and they are often applied without the perceivers’ intention or conscious awareness (Macrae & Bodenhausen 2000). This last point relates to the third aspect of person schemata. Person schemata are unities of characteristic features integrated at the level of persons. All these singular features are integrated into person models that enable us to develop detailed and extensive expectations of behaviour.

(iii) *Integration of characteristic features at the level of perceiving the whole person:* Although I have presented evidence that some single features are very salient for transferring social information, there is also much evidence that these features are normally combined with a variety of others to form an integrated impres-

¹⁴ Interestingly, Germans could perceive the friendliness of students from the US and UAE partially (as well as the other way around), while students from UAE and USA could not read the level of friendliness from the other culture at all (Bente et al. 2010).

¹³ We leave the question open as to what extent person schemata are constituted by innate or by learned dispositions. The examples mentioned above indicate that they involve properties of both kinds.

sion of a person that I call a person schema. We have seen evidence for the key role of gaze detection in registering another person's direction of attention (see ii). But there is further evidence that gaze alone is not the critical source of information; we actually seem to rely on an integrated evaluation on the basis of perceiving gaze, head, and body position (Frischen et al. 2007). The same holds for evaluation of the basic features sex, race, and age. Although isolated facial features are often sufficient to determine a person's sex, research has indicated that sex categorization is based on the integration of several features (Baudoin & Humphreys 2006; Bruce et al. 1993; Brown & Perrett 1993; Roberts & Bruce 1988; Schyns et al. 2002). Concerning face, the best available theory of face recognition seems to be Haxby's account (Haxby et al. 2000), according to which there are two distinguishable processes, one leading to face identification by focussing more on invariant core features, and the other leading to registering facial expression by relying on varying features. Furthermore, there is evidence that there are two different neural circuits for face perception and body perception (see the review by Macrae & Quadflieg 2010), both playing a core role in registering face or body identity, and playing an extended role in registering face or body expression in a given situation. And the integration processes are not limited to this level (Martin & Macrae 2007). Since we know that information about facial and bodily features is integrated, e.g., in the evaluation of emotional expression, we can therefore characterize a sequence of integration processes as leading finally to a person impression in a situation, which may be stored as a person schema in memory.

5.2 A model of forming a person schema

How can we best describe this process of forming a person schema? In general terms, the same complex process takes place in the case of perceiving a person and forming a person impression in a given situation as takes place when we perceive an object. I describe the process according to the model of object

perception developed by Ernst & Bühlhoff (2004), and I have already shown in detail that it can do justice to our recognition of emotions (Newen et al. forthcoming). The overall process comprises bottom-up processes starting with basic visual features that are modulated either by feature combination (if two features provide complementary information), or by feature integration. The latter can be modelled as a Bayesian weighting process that leads to the most probable intermediate estimate given the input. Further integration processes then lead from the most probable estimate to a stable percept of an object in the case of object perception, and to a stable person impression in the case of person perception. This model explicitly accounts not only for bottom-up but also for top-down processes, in the form of so-called cognitive penetration. I have sketched a plausible but in no way complete model of the formation of a person impression (see figure below). According to the evidence I have presented so far, it is plausible to suggest that at the level of intermediate estimates in the process of forming an impression of a person, we find (a) an estimation of a core person identity, (b) an estimate of situational emotions, intentions, and actions, as well as (c) an estimation of social status, person abilities, and individual personality traits. An important step in the model is the association of visual features with socially-anchored stereotypes (see above) which allows us to develop rich intermediate estimates, e.g., of the other's emotional situation, social status, etc.

Numerous lines of research (Albright, Kenny, & Malloy, 1988; Ambady & Rosenthal, 1992; Behling & Williams, 1991; Borkenau & Liebler, 1992; Kenny, Horner, Kashy, & Chu, 1992; Norman & Goldberg, 1966; Secord, Dukes, & Bevan, 1954) have provided compelling evidence that trait evaluations are readily drawn from a person's physiognomy (i.e., facial features), outer appearance (i.e., clothing), or demeanor (i.e., posture, walking, style). (Macrae & Quadflieg 2010, p. 433)

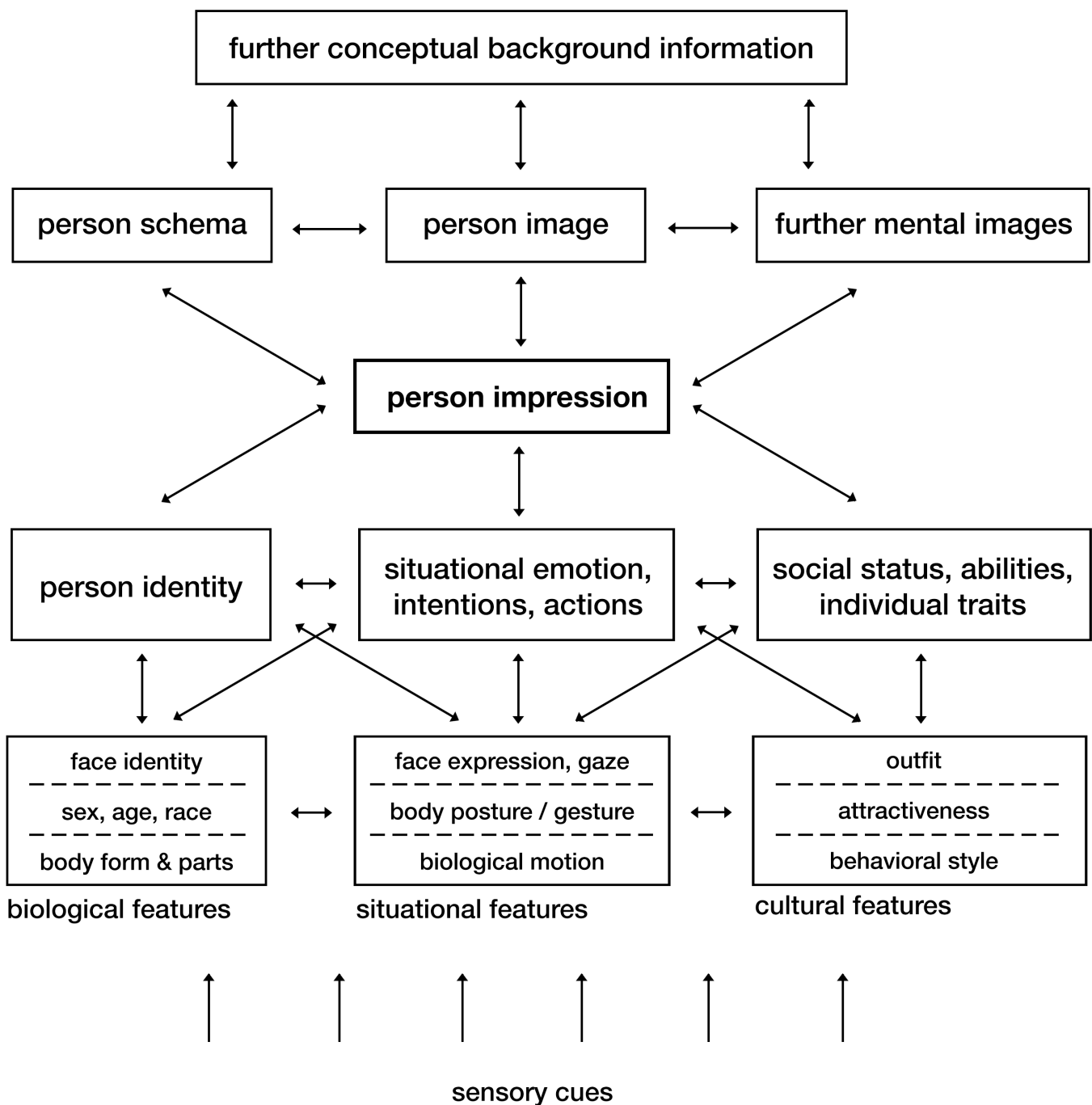


Figure 1: A model of the dynamics of bottom-up and top-down processes leading to a stable person impression by relying on person images and/or person schemata

Finally, I highlight that the top-down processes are able to interfere in this process of combination and integration very early in the visual information processes: for example, it has been shown that the activation of a race concept on the basis of the form of a face (African versus European face format) changes the perception of colour in the face, while colour is known to be represented in V4 as part of early visual

brain processes. The same hue of colour is seen as more dark in the African face than in the European face (Levin & Banaji 2006). Thus we have to admit that the process of feature-combination and integration is highly dynamic, involving simultaneous activation of features rooted in bottom-up and top-down processes, finally reaching the most probable and usually stable person impression. The dynamic is de-

scribed in detail for the case of object perception in Vetter & Newen (2014); it is postulated for person categorization in Macrae & Martin (2007), and analysed according to the levels of processing that lead to person construal in Freeman & Ambady (2011). Figure 1 is a sketch of the formation of a person impression according to my account.

A person schema emerges as the result of direct perception of a person, where this may be either basic or relatively smart perception; yet it usually remains implicit, and is not amenable to linguistic description. A typical example of person schema based on basic perception is the everyday experience of seeing a person only briefly in a single situation, whereupon it is difficult for us to describe the person—particularly her face. While we can often easily recognize the person, it may take hours with a professional to end up with an adequate “identikit” picture such as those produced at police stations. A person schema based on smart perception might be, for instance, a person schema that includes a lot of top-down activation—for example, while on campus, perhaps I see a person of typical student age dressed like a law student, and thus activate the “rich person” schema that is the basis for my everyday smooth interaction with law students, and which differs (despite overlaps) from my person schema for students in natural sciences. If we not only develop implicit practical knowledge regarding our use of the person impression (independent from its richness), but also develop explicit *knowledge* pertaining to the relevant person information, or at least develop easy explicit access to it, then we go beyond a person schema. We can characterize this new unified information as a person image.

5.3 Person images

In detail, then, what is a person image? A person image is a unity of relatively easily and explicitly available information about a person, including her mind-set. On the basis of typically implicit person schemata, young children learn to develop explicit *person images*. These are models of individual subjects or groups. In the

case of individual subjects, they may include names, descriptions, stories, whole biographies, and visual images highlighting both mental and physical dispositions as well as episodes. Person images are essentially developed not only by observation but also by telling, exchanging, and creating stories (or “narratives”).¹⁵ Person images presuppose the capacity to explicitly distinguish the representation of my own mental and physical phenomena from the representation of someone else’s mental and physical phenomena. This ability develops gradually, reaching a major and important stage when children acquire the so-called explicit theory-of-mind ability (operationalized by the false-belief task, see Wimmer & Perner 1983).¹⁶ Then they are able to construct explicit person images by characterizing a person such that they attribute a biography to an individual. There is strong folk-psychological evidence that we have explicit person models of the people we deal with extensively, e.g., family members, and people about whom we tend to have a lot of explicit knowledge. The same is true for relevant groups of persons we deal with often. Even in professional contexts this leads to judgments that can be inadequate: the apparent association between wearing revealing clothes and immodesty and promiscuity has been shown to cause not only laypeople but also police officers and judges to hold victims of rape to be responsible for their having been assaulted (Lennon et al. 1999). An essential part of becoming an adult is learning to interact socially with other humans, by developing sophisticated and explicit person images of the groups of professions we have to come to any sort of arrangement with. We often have explicit beliefs about medical doctors, managers, secretaries, craftspeople, etc., and we try to deploy these beliefs to deal with these people in a smooth and efficient way. When we

¹⁵ This is the aspect of the narrative approach to understanding other minds, mentioned above (e.g., Hutto 2008). But narratives are only one method of establishing a person model. Representatives of a pure narrative approach underestimate the importance of other sources, such as perceptions, feelings, interactions, etc., which often do not involve narratives.

¹⁶ There is a long and not fully understood process of development from implicit false belief sensitivity to explicit false belief understanding (de Bruin & Newen 2012a; 2012b). Person images actually presuppose an explicit representation of false beliefs.

have stored a person image in memory, and are placed in a new situation in which we see and recognize the person, there is evidence that we immediately activate the biographical knowledge we have available. For example, when test persons were asked to judge the traits of target individuals from photographs, the test persons' responses continue to be influenced by what they have explicitly learned about the people in question (Uleman et al. 2005). A recent neuroimaging study (Hassabis et al. 2013) indicated that when test persons were asked to predict the behaviour of persons, they essentially relied on prior knowledge of personality traits, which in this particular study were implemented in two ways, namely as agreeableness (the tendency toward altruism, cooperation, and the valuing of harmony in interpersonal relationships as opposed to antisocial and exploitative behaviours) and as extroversion (in contrast to introversion). The test person became acquainted with four types of personalities that had been constructed from combinations of high and low versions of agreeableness, on the one hand, and high and low versions of extroversion, on the other. In the test situation they had to predict the behaviour of four specific persons who were exemplars of the four personality types. The authors report that the predictions of behaviour were mainly based on personality traits and that the latter also had rather clear neural correlates: by using functional magnetic resonance imaging (fMRI) the authors showed that there is a neural correlate for recognizing (and imagining) high agreeableness (in contrast to low), namely in the left LTC (lateral temporal cortex) and dorsal mPFC (medial prefrontal cortex), as well as for recognizing (and imagining) high extroversion (in contrast to low), namely in the pCC (posterior cingulate cortex); in addition the recognition (and imagination) of one of the four personality types was correlated with four distinctive patterns in the anterior medial prefrontal cortex (mPFC). In line with my proposal, the authors of the fMRI study write: "Different patterns of activation in the anterior mPFC could reliably distinguish between the different people whose behavior was being imagined. It is hypothesized that this

region is responsible for assembling and updating personality models" (Hassabis et al. 2013). Since the study was based on explicit evaluation of personality features or types, I take this to support the existence of person images. Yet even if the reader accepts the idea of person models, she may be sceptical about whether we need to distinguish person *schemata* and person *images*.

5.4 Why should we distinguish person schemata and person images?

A very convincing case that forces us to make a distinction between person schemata and person images comes from taking a closer look at a typical patient suffering from Capgras syndrome, a misidentification syndrome. Sufferers have the delusional belief that one of their closest relatives, e.g., their wife, has been replaced by an impostor. Such a patient typically says things like "this person looks exactly like my wife, she even speaks and behaves like my wife and she expresses her typical desires but she is not my wife" (Davies et al. 2001); thus, one aspect of this mental disorder is the observation that all the features explicitly believed to be possessed by the wife are correctly attributed. We can account for this by asserting that the patient has an intact person image of his wife. Nevertheless, the usual person identification has gone wrong. According to a standard analysis, what is lacking in the case of the Capgras patient is a feeling of familiarity that normally comes with perceiving a well-known person. How can we account for this in the new framework? When perceiving his wife, the subject intuitively develops and activates a person schema. One aspect of the person schema is the person's identity.¹⁷ As the Capgras case nicely illustrates, the registration of a person's identity is a result of an integration process that relies not only on visual features but also on an implicit emotional evaluation, and that these together trigger an explicit judgment. While the

¹⁷ The involvement of identity already at the level of implicit schemata is supported by Haxby's model of face perception according to which we have to distinguish a core cognitive system involving the recognition of face identity and an extended cognitive system which is enabling the recognition of facial expression (Haxby et al. 2000).

visual recognition fits, here the emotional evaluation is inadequate and the feeling of familiarity is lacking; and in the case of this disorder, the Bayesian integration process for these features leads to an implausible result, since the emotional mistake overrides the visual adequateness. Thus, the Capgras patient has an adequate person image of his wife but an incorrect person schema, and the tension between the two is solved by developing the (implausible) hypothesis that she is an imposter. This analysis is in line with two-factor theories of the Capgras disorder, according to which two distinct factors cause the phenomenon¹⁸: first, the lack of familiarity, and, second, a local breakdown of rationality that enables the irrational belief-formation on the basis of a severely disturbed person schema (Davies et al. 2001).¹⁹ Several other cases seem to be accounted for if we accept the evidence for a two-factor theory of person modelling—namely a first level of intuitive and implicit person impression and a second level of explicit person evaluation, which are described respectively as intuitive person schemata and explicit person images.

A contrast case to Capgras syndrome is the Fregoli syndrome, wherein a patient has the delusional belief that one and the same person, usually a persecutor, is following her, who is able to radically change his outer appearance. The sufferer then connects people with rather different outer appearances and treats them as the same persecutor. One explanation, still in need of testing, is that this time the feeling of familiarity is developed too often, probably by top-down initiation due to the delusional belief

that the subject is being persecuted. The delusional belief, together with an inadequate feeling of familiarity, may explain the syndrome.²⁰ But again we need to distinguish the two factors: a level of implicit feeling or impression, and a level of explicit judgment. This time the delusion produces a breakdown of rational judgment formation, i.e., the person model of the other is strongly influenced by the delusion: the person schema formation may be largely intact but has a local defect due to being dominated by the delusional belief. In general, monothematic delusions (delusions about a single belief content) seem to rely on two factors (Coltheart et al. 2007): “[o]ne factor has to explain the strange experiences patients claim to have, while the other factor has to explain the misattribution of actions and thoughts” (Vosgerau & Newen 2007, p. 40).

Are there nonpathological everyday cases that support the distinction between person schema and person image? One illustration can be drawn from Mark Twain’s “Huckleberry Finn.” At first Huck helps the slave Jim to escape from slavery; but then he rethinks his support in the light of the law, and forms the judgment that he should turn him in to the slavehunters. But when he has the opportunity to do so, Huck actually ends up protecting Jim. Why does he do this? Huck has a person schema of Jim that is constituted by a person impression according to personal interactions that are dominated by empathy; thus he has a positive impression of Jim and there exists between them a growing friendship. On the other hand, he has a person image of Jim that is dominated by the fact that he is a slave, such that he has to accept his role in society, to do the hard work, to live without freedom, and thus that it is forbidden to aid his escape. Cases of tension between an intuitive person impression (being helpful, being peaceful) and a person image dominated by the knowledge that the same person is a pathological murderer are often reported by judges and policemen. A less dramatic tension seems to be part of our everyday experience of “false” friends (we may still think of someone as

¹⁸ In the literature there are discussed one-factor accounts to explain mental disorders, e.g., in the case of schizophrenia (Gallagher 2004): a top-down approach argues that disturbances of higher-order cognition is the only source for thought insertion (Stephens & Graham 2000) while a bottom-up approach argues that thought insertion is a product of disturbances of neural or basic cognitive processes (like perception). Most of the recent accounts are hybrid account which we call two-factor theories.

¹⁹ The fact that person identity as a component of person schema formation is not only based on visual but also on an emotional evaluation is supported by the case of prosopagnosia, i.e., the inability to recognize the face of the person one is seeing, even though one is able to see and perceive the rest of the person adequately. Despite the fact that a person suffering from prosopagnosia is not able to see the familiarity of the face, we can measure increased skin conductance for familiar but not unfamiliar faces, thereby demonstrating intact (albeit covert) emotional recognition of known others (de Haan et al. 1992; Tranel & Damasio 1985).

²⁰ For a discussion of delusional phenomena, see Coltheart et al. (2007) and Hirstein (2005).

a friend while implicitly already noticing signs of unfair treatment), though of course the tension can also exist the other way around. As illustrated above, the visual features of a person are often loaded with social information, and often involve the activation of negative prejudices which, after a more careful investigation of the person, can be opposed by a positive person image. The general functional role of person models is to simplify the structuring and evaluation of social situations, to enable a quick evaluation of the person in a given situation, and to initiate adequate behaviour. An additional special functional role of person models consists in stabilizing my self-estimation, since there is a strong tendency to have positive stereotypes of one's own in-group members and negative stereotypes of the out-groups' (see Volz 2008, p. 19). These examples illustrate not only that we need to distinguish the person schema and person image, but also that we have a tendency towards harmonizing both. Thus, if one of them is disturbed we tend to adjust the other, which may result not only in wrong judgments about persons, but in extreme cases may become an aspect of a mental disorder, as described above. Finally, to distinguish them is compatible with the claim that a person image may often gradually evolve on the basis of a person schema such that partially the same information about a person changes the status of accessibility from implicit to explicit. But we also have to distinguish both kinds of person models because often an implicit representation of a person as unfriendly exists simultaneously with an explicit evaluation of the same person as friendly.

5.5 Person model theory (PMT) and its relation to other main theories

The central claim of PMT is that we organize information about others by forming person models. We account for a multiplicity of epistemic access strategies, while direct perception and interaction are the main source for person schema formation. Person image formation is based on all the epistemic strategies we have examined, including theory-based inferences and (high-level) simulation strategies. Why, then, is

PMT not a version of TT? Person models are more general and allow for a unification of rather parsimonious information about a person, which does not warrant being called a theory since it does not form even a minimal package of systematically-interconnected beliefs. As we learn more and more about the same person, our person model may develop into a theory. Thus, this is not to deny that we often have rich person models that are theories; and thus I can account for the empirical evidence that supporters of TT tend to rely on. A further question concerns how PMT is related to ST. Simulation is one epistemic strategy in which person models are used to understand others: if I have evidence that another person is similar to me in relevant respects, then I may use my self-model, either the self-schema or the self-image, to produce an explanation or a prediction of the other's behaviour. But I also often have clear knowledge that the other is different from me in relevant respects, especially when there are great differences in the three main categories—sex, age, and race—or in cultural background. In such cases simulation is not used. Although simulation is a worthy epistemic strategy, it is only of limited and constrained use in everyday understanding. How is PMT related to interaction theory and direct perception theories? It explicitly accepts the important role of both as epistemic strategies, but insists that in addition to understanding others in situations of direct interaction there is also often an understanding of others just by observation. The use of these two strategies seems to depend heavily on the personality traits of the person who aims to understand another: while extroverts mainly rely on interaction, introverts (who avoid social contact) mainly rely on observation. Furthermore, these theories do not offer an answer to the main question addressed in this article, namely how we organize the information about other people that we already have. The narrative account offers one answer here, and again we can account for the role of narratives that in the case of rather rich person models may be sources for creating or enriching the models further, or they may also concern the way a person model is memorized. But the narrative account

person model theory

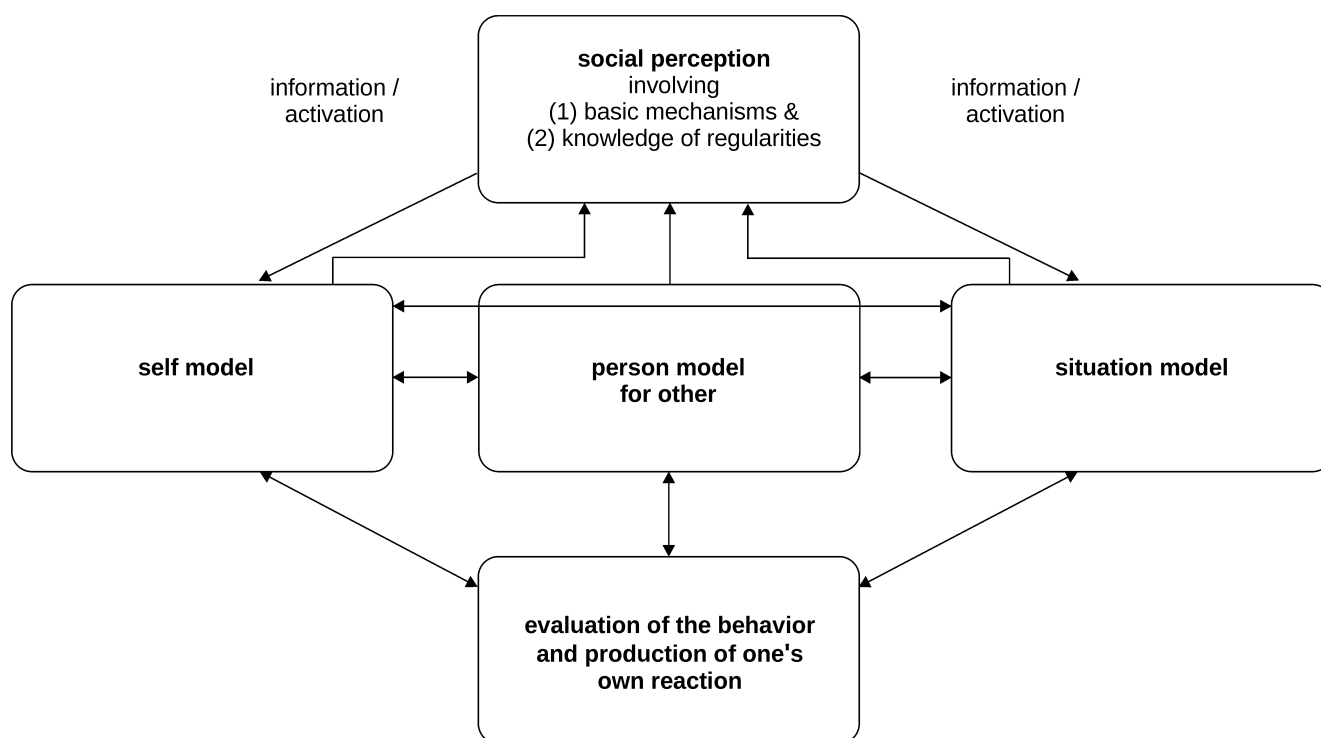


Figure 2: Interaction of person models with situation models in understanding others

alone ignores the strong relevance of our intuitive understanding of others as it is anchored in person schemata. This short overview, then, indicates that all of the evidence that representatives of other theories put forward can be integrated into this view, while there is further evidence for my theory, e.g., rich evidence that there is an integration of information into person models by person perception. Notably, PMT allows us to account for certain mental disorders, and I have cited evidence from a very recent fMRI study that is further supportive of the organization of information according to person models.

5.6 Widening PMT: Person models, situation models and culture

Does PMT give us the complete story about understanding others? What about my understanding of a person whom I only see from behind, when queuing at a self-service restaurant? Here it seems sufficient to predict her behaviour just by expecting her to act according to the so-

cial conventions of a self-service restaurant. Understanding the situation alone seems to be sufficient for an understanding of and interaction with the other.²¹ This is an important observation that suggests a widening of my theory: we do not only create person models, but also situation models, and our understanding of others uses both types of model as input and selects the most helpful model for evaluating the other person. If I have no person model of this individual, if seeing someone from the back gives me only very parsimonious information, and if I am only interested in getting my lunch, then the situation model may be dominant in dealing with persons in this context. As soon as minimal enrichment of person information is available we naturally tend to rely on person models. The fact that situation models are used at all is supported by successful artificial intelligence (AI) studies working with scripts and

²¹ These types of cases are considered in Gallagher & Hutto (2008), in the section “Pragmatic Intersubjectivity”. Their view is close to a multiplicity view. A minor criticism is that we have to account for such cases independently from being in interaction with someone. They may also involve only observing the other.

frames that can account for human behaviour (Schank & Abelson 1977). Furthermore, in Asian cultures the understanding of other people seems to rely much more on social conventions, since people are strongly expected to behave according to these conventions. In general, situation models are more important for understanding others in “collectivistic” cultures than in individualistic cultures where explanations and predictions of behaviour are usually more reliant on individual belief–desire explanations. Such observations as these require us to give an account of situation models. This can be easily done by widening the theory of understanding others such that it includes situation models, as well as the interdependence of personal models and situation models. It can also include a dynamic, involving bottom-up and top-down processes that lead to an activation or construction of the most plausible person model for interacting with, explaining, or predicting the behaviour of the other person.²² Here is a rough outline of the process leading to understanding others in the rich sense of interacting, such as in observing, explaining, or predicting (see Figure 2).

In general, we should note the important role of culture in shaping our way of modelling persons (Vogeley & Roepstorff 2009). As we have seen, culture modulates the relevance of person models in relation to situation models. But it also influences our formation of person models, for example by shaping our person perception. To illustrate: Japanese individuals are encouraged to be sociable and cooperative (Moskowitz et al. 1994), to be affiliative rather than competitive (Yamaguchi et al. 1995), and to show obligation to others (Oyserman et al. 1998). Concerning dominance and subordination, Japanese people learn to be rewarded for subordinate behaviour, while Americans learn to be rewarded for dominant behaviour. This

also shapes the perception of dominance and subordination in others. Typical neurological activations of the mesolimbic reward system can be shown to be shaped by the respective culture: Americans show a higher activation of this system when doing and seeing dominant behaviour (in contrast to subordinate behaviour) while with Japanese people we can observe the opposite: they show a higher activation of exactly the same system when doing and seeing subordinate behaviour (Freeman et al. 2009). Thus, the perception of dominant and subordinate behaviour is connected with opposite evaluations (Americans highly esteem dominance while Japanese people highly esteem subordinate behaviour) and a different set of personality traits. Cultural influences on the psychological and neural level are also reported for self-models: on the psychological level, the difference between an Asian interdependent self and a Western independent self was reported by Markus & Kitayama (1991), while a respective difference in neural correlates was also recently discovered (Sui & Han 2007).

6 Conclusion

Our understanding of other minds is based epistemically on a multiplicity of strategies, the core strategies being direct perception, interaction, simulation, and theory-based inferences (including learning from narratives). The most important aspect of understanding others is the activation of prior knowledge of individuals or groups of persons. This is organized into person models. The main claim of PMT is that we rely on *person models* to understand others. These person models form the basis for perceiving and evaluating persons, their social behaviour, and their mind-set. We develop person models for ourselves, for other individuals, and for groups of persons (group models). Furthermore, all types of person models can be realized on two levels: (implicit) person schemata and (explicit) person images. A *person schema* is a bundle of information including information about sensory-motor abilities, voice, face, basic mental dispositions, etc., and such schemata are intuitively used, implicitly developed, and not usually

²² There is already one dynamic model of person construal available in the literature that also supports my dynamic theory of understanding others with person models, i.e., the model of Freeman & Ambady (2011). Despite its merits in describing social perception in more detail as regards the interrelation of bottom-up and top-down processes, the authors neither account for the claim that our rich prior information is mainly organized on the level of persons (not faces or subpersonal features), nor do they account for the interaction between person models and situation models.

easily accessible for linguistic report. A *person image* is a unity of explicitly-registered mental and physical dispositions as well as situational features (like perceptions, emotions, attitudes, etc.) that is usually easily accessible for linguistic report (albeit sometimes with the help of gesture, drawings, etc.). The PMT has several advantages over existing accounts of social understanding (e.g., TT, ST, and interaction theory), since it can account for all of the following criteria:

1. It explains specific and more general social understanding of particular individuals in terms of individual person models and group person models. (Not accounted for in ST.)
2. It accounts for the difference, for which evidence is presented, between implicit, intuitive forms of social understanding and explicit deliberative ones by appealing to the role of person schemata and person images respectively. (Not accounted for in interaction theory.)
3. It does justice to folk-psychological evidence that we understand very familiar persons much better than unfamiliar ones: We have rich person images of individuals with whom we are very familiar. (Deficit of all former theories.)
4. It marks adequately in what ways our understanding of others and our self-understanding are interdependent, e.g., in special cases of simulation, understanding the other relies on self-models. (Generally not accounted for in TT.)
5. It offers an adequate framework that is in line with the best explanations of some mental diseases in understanding others, such as the Capgras and Fregoli syndromes. (Deficit of ST.)
6. It can account for cultural differences in social understanding: Future research will show how person models vary with culture, and we have already illustrated that it varies in the case of self-models between Asian and Western cultures. (Not accounted for in any former theory.)

Thus, PMT is at least a serious alternative account, and certainly a candidate for future investigation.

Acknowledgements

I wish to express special thanks to Luca Barlassina, Kai Vogeley, Anna Welpinghus, and Tobias Starzak for their helpful comments. This paper is part of the project “Social Information Processing and Culture” funded by the VolkswagenStiftung.

References

- Adams, R. B., Jr. & Kleck, R. E. (2003). Perceived gaze direction and the processing of facial displays of emotion. *Psychological Science*, 14 (6), 644-647.
- Albright, L., Kenny, D. A. & Malloy, T. E. (1988). Consensus in personality judgments at zero acquaintance. *Journal of Personality and Social Psychology*, 55 (3), 387-395. [10.1037/0022-3514.55.3.387](#)
- Ambady, N. & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111 (2), 256-274. [10.1037/0033-2909.111.2.256](#)
- Bar, M., Neta, M. & Linz, H. (2006). Very first impressions. *Emotion*, 6 (2), 269-278. [10.1037/1528-3542.6.2.269](#)
- Baron-Cohen, S. (1995). *Mindblindness. An essay on autism and theory of mind*. Cambridge, MA: MIT Press.
- Baudoin, J.-Y. & Humphreys, G. W. (2006). Configural information in gender categorisation. *Perception*, 35 (4), 431-450. [10.1068/p3403](#)
- Behling, D. U. & Williams, E. A. (1991). Influence of dress on perception of intelligence and expectations of scholastic achievement. *Clothing and Textiles Research Journal*, 9 (4), 1-7. [10.1177/0887302X9100900401](#)
- Bente, G., Leuschner, H., Al Issa, A. & Blascovich, J. J. (2010). The others: Universals and cultural specificities in the perception of status and dominance from non-verbal behavior. *Consciousness and Cognition*, 19 (3), 762-777. [10.1016/j.concog.2010.06.006](#)
- Bentin, S. & Deouell, L. (2000). Structural encoding and identification in face processing: ERP evidence for separate mechanism. *Cognitive Neuropsychology*, 17 (1-3), 35-54.
- Bentin, S., Allison, T., Puce, A., Perez, E. & McCarthy, G. (1996). Electrophysiological studies of face perception in humans. *Journal of Cognitive Neuroscience*, 8 (6), 551-565. [10.1162/jocn.1996.8.6.551](#)
- Bertin, E. & Striano, T. (2006). The still-face response in newborn, 1.5-, and 3-month-old infants. *Infant Behavior and Development*, 29 (2), 294-297. [10.1016/j.infbeh.2005.12.003](#)
- Borkenau, P. & Liebler, A. (1992). Trait inferences: Sources of validity at zero acquaintance. *Journal of Personality and Social Psychology*, 62 (4), 645-657. [10.1037/0022-3514.62.4.645](#)
- Brewer, M. B. (1988). A dual-process model of impression formation. In R. S., Jr. Wyer & T. K. Srull (Eds.) *Advances in Social Cognition* (pp. 1-36). Mahwah, NJ: Erlbaum.
- Brown, E. & Perrett, D. I. (1993). What gives a face its gender? *Perception*, 22 (7), 829-840. [10.1068/p220829](#)
- Bruce, V., Burton, A. M., Hanna, E., Healey, P., Mason, O., Coombes, A., Fright, R. & Linney, A. (1993). Sex discrimination: How do we tell the difference between male and female faces? *Perception*, 22 (2), 131-152. [10.1068/p220131](#)
- Buckley, S. J., Bird, G. & Sacks, B. (2002). Social development for individuals with down syndrome: An overview.
- Calder, A. J., Keane, J., Manes, F., Antoun, N. & Young, A. W. (2000). Impaired recognition and experience of disgust following brain injury. *Nature Neuroscience*, 3, 1077-1078. [10.1038/80586](#)
- Carroll, J. M. & Russell, J. A. (1996). Do facial expressions signal specific emotions? Judging emotion from the face in context. *Journal of Personality and Social Psychology*, 70 (2), 205-218. [10.1037/0022-3514.70.2.205](#)
- Carruthers, P. (2009). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences*, 32 (2), 121-182. [10.1017/S0140525X09000545](#)
- Cleveland, A. & Striano, T. (2007). The effects of joint attention on object processing in 4- and 9-month-old infants. *Infant Behavior and Development*, 30 (3), 499-504. [10.1016/j.infbeh.2006.10.009](#)
- Coltheart, M., Langdon, R. & McKay, R. (2007). Schizophrenia and monothematic delusions. *Schizophrenia Bulletin*, 33 (3), 642-647. [10.1093/schbul/sbm017](#)
- Davies, M., Coltheart, M., Langdon, R. & Breen, N. (2001). Monothematic delusions: Towards a two-factor account. *Philosophy, Psychiatry and Psychology*, 8 (2/3), 133-158. doi: [10.1353/ppp.2001.0007](#)
- de Bruin, L., van Elk, M. & Newen, A. (2012). Reconceptualizing second-person interaction. *Frontiers in Neuroscience*, 151, 1-10. [10.3389/fnhum.2012.00151](#)
- de Bruin, L. & Newen, A. (2012a). An association account of false belief understanding. *Cognition*, 123 (2), 240-259. [10.1016/j.cognition.2011.12.016](#)
- (2012b). The developmental paradox of false belief understanding: A dual-system solution. *Synthese*, 191 (3), 297-320. [10.1007/s11229-012-0127-6](#)
- de Haan, E. H. F., Bauer, R. M. & Greve, K. W. (1992). Behavioural and physiological evidence for covert face recognition in a prosopagnosic patient. *Cortex*, 28 (1), 77-95. [10.1016/S0010-9452\(13\)80167-0](#)
- Devine, P. G. & Elliot, A. J. (1995). Are racial stereotypes really fading? The Princeton trilogy revisited.

- Personality and Social Psychology Bulletin*, 21 (11), 1139-1150. [10.1177/01461672952111002](https://doi.org/10.1177/01461672952111002)
- Dion, K., Berscheid, E. & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology*, 24 (3), 285-290. [10.1037/h0033731](https://doi.org/10.1037/h0033731)
- Eagly, A. H., Ashmore, R. D., Makhijani, M. G. & Longo, L. C. (1991). What is beautiful is good, but ...: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin*, 110 (1), 109-128. [10.1037/0033-2909.110.1.109](https://doi.org/10.1037/0033-2909.110.1.109)
- Eimer, M. & Holmes, A. (2002). An ERP study on the time course of emotional face processing. *NeuroReport*, 13 (4), 427-431. [10.1097/00001756-200203250-00013](https://doi.org/10.1097/00001756-200203250-00013)
- Eimer, M. & McCarthy, R. A. (1999). Prosopagnosia and structural encoding of faces: Evidence from event-related potentials. *NeuroReport*, 10 (2), 255-259. [10.1097/00001756-199902050-00010](https://doi.org/10.1097/00001756-199902050-00010)
- Ekman, P. (1972). Universals and cultural differences in facial expressions of emotions. In J. Cole (Ed.) *Nebraska Symposium on Motivation, 1971, Vol. 19* (pp. 207-283). Lincoln, NE: University of Nebraska Press.
- (1999). Basic emotions. In T. Dalgleish & M. J. Power (Eds.) *The Handbook of Cognition and Emotion* (pp. 45-60). New York, NY: Wiley.
- Ekman, P., Friesen, W. V. & Ellsworth, P. (1972). *Emotion in the Human Face*. New York, NY: Pergamo.
- Ernst, M. O. & Bühlhoff, H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences*, 8 (4), 162-168. [10.1016/j.tics.2004.02.002](https://doi.org/10.1016/j.tics.2004.02.002)
- Fiebach, A. & Coltheart, M. (under review). *Various ways to understand other minds*.
- Fiske, S. T. & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in Experimental Social Psychology*, 23, 1-74. [10.1016/S0065-2601\(08\)60317-2](https://doi.org/10.1016/S0065-2601(08)60317-2)
- Freeman, J. B. & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological Review*, 118 (2), 247-279. [10.1037/a0022327](https://doi.org/10.1037/a0022327)
- Freeman, J. B., Rule, N. O. & Ambady, N. (2009). The cultural neuroscience of person perception. *Progress in Brain Research*, 178, 191-201. [10.1016/S0079-6123\(09\)17813-5](https://doi.org/10.1016/S0079-6123(09)17813-5)
- Frey, S. (1999). *Die nonverbale Kommunikation*. Bern, SUI: Huber.
- Frischen, A., Bayliss, A. P. & Tipper, S. P. (2007). Gaze cueing of attention: Visual attention, social cognition, and individual differences. *Psychological Bulletin*, 133 (4), 694-724. [10.1037/0033-2909.133.4.694](https://doi.org/10.1037/0033-2909.133.4.694)
- Gallagher, S. (2001). The practice of mind: Theory, simulation, or interaction? *Journal of Consciousness Studies*, 8 (5-7), 83-107.
- (2004). Neurocognitive models of schizophrenia: A neurophenomenological critique. *Psychopathology*, 37 (1), 8-19. [10.1159/000077014](https://doi.org/10.1159/000077014)
- (2005). *How the body shapes the mind*. Oxford, UK: Oxford University Press.
- (2007). Simulation trouble. *Social Neuroscience*, 2 (3), 353-365. [10.1080/17470910601183549](https://doi.org/10.1080/17470910601183549)
- (2008). Direct perception in the intersubjective context. *Consciousness and Cognition*, 17 (2), 535-543. [10.1016/j.concog.2008.03.003](https://doi.org/10.1016/j.concog.2008.03.003)
- Gallagher, S. & Hutto, D. (2008). Understanding others through primary interaction and narrative practice. In J. Zlatev, T. P. Racine, C. Sinha & E. Itkonen (Eds.) *The shared mind: Perspectives on intersubjectivity* (pp. 17-38). Amsterdam, NL: John Benjamins.
- Gliga, T. & Dehaene-Lambertz, G. (2005). Structural encoding of body and face in human infants and adults. *Journal of Cognitive Neuroscience*, 17 (8), 1328-1340. [10.1162/0898929055002481](https://doi.org/10.1162/0898929055002481)
- Gobet, F. (1997). Roles of pattern recognition and search in expert problem solving. *Thinking and Reasoning*, 3 (4), 291-313. [10.1080/135467897394301](https://doi.org/10.1080/135467897394301)
- Gobet, F. & Simon, H. A. (1996). Recall of rapidly presented random chess positions is a function of skill. *Psychonomic Bulletin & Review*, 3 (2), 159-163. [10.3758/BF03212414](https://doi.org/10.3758/BF03212414)
- Goldman, A. I. (2006). *Simulating minds. The philosophy, psychology, and neuroscience of mindreading*. Oxford, UK: Oxford University Press.
- Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences*, 16 (1), 1-14. [10.1017/S0140525X00028636](https://doi.org/10.1017/S0140525X00028636)
- Gopnik, A. & Meltzoff, A. (1997). *Words, thoughts, and theories*. Cambridge, MA: MIT Press.
- Griffiths, P. E. (1997). *What Emotions Really Are. The Problem of Psychological Categories*. Chicago, IL: Chicago University Press.
- Hassabis, D., Spreng, R. N., Rusu, A. A., Robbins, C. A., Mar, R. A. & Schacter, D. L. (2013). Imagine all the people: how the brain creates and uses personality models to predict behavior. *Cerebral Cortex*, March 5. [10.1093/cercor/bht042](https://doi.org/10.1093/cercor/bht042)
- Haxby, J. V., Hoffman, E. A. & Gobbini, M. A. (2000). The distributed human neural system for face percep-

- tion. *Trends in Cognitive Sciences*, 4 (6), 223-233.
[10.1016/S1364-6613\(00\)01482-0](#)
- Heberlein, A. S., Adolphs, R., Tranel, D. & Damasio, H. (2004). Cortical regions for judgments of emotions and personality traits from pointlight walkers. *Journal of Cognitive Neuroscience*, 16 (7), 1143-1158.
[10.1162/0898929041920423](#)
- Herrmann, M. J., Ehlis, A. C., Muehlberger, A. & Fallgatter, A. J. (2005). Source localization of early stages of face processing. *Brain Topography*, 18 (2), 77-85. [10.1007/s10548-005-0277-7](#)
- Hirstein, W. (2005). *Brain fiction. Self deception and the riddle of confabulation*. Cambridge, MA: MIT Press.
- Holmes, A., Vuilleumier, P. & Eimer, M. (2003). The processing of emotional facial expression is gated by spatial attention: Evidence from event-related brain potentials. *Cognitive Brain Research*, 16 (2), 174-184.
[10.1016/S0926-6410\(02\)00268-9](#)
- Hutto, D. (2008). *Folk-psychological narratives*. Cambridge, MA: MIT Press.
- Itier, R. J. & Taylor, M. J. (2004). N170 or N1? Spatiotemporal differences between object and face processing using ERPs. *Cerebral Cortex*, 14 (2), 132-142.
[10.1093/cercor/bhg111](#)
- Kenny, D. A., Horner, C., Kashy, D. A. & Chu, L. (1992). Consensus at zero acquaintance: Replication, behavioral cues, and stability. *Journal of Personality and Social Psychology*, 62 (1), 88-97.
[10.1037/0022-3514.62.1.88](#)
- Kuzmanovic, B., Schilbach, L., Lehnhardt, F. G., Bente, G. & Vogeley, K. (2011). A matter of words: Impression formation in complex situations relies on verbal more than on nonverbal information in high-functioning autism. *Research in Autism Spectrum Disorders*, 5, 604-613.
- Lennon, S. J., Johnson, K. K. P. & Schulz, T. L. (1999). Forging linkages between dress and law in the U.S., part I: Rape and sexual harassment. *Clothing and Textiles Research Journal*, 17 (3), 144-156.
[10.1177/0887302X9901700305](#)
- Levin, D. T. & Banaji, R. (2006). Distortions in the perceived lightness of faces: The role of race categories. *Journal of Experimental Psychology: General*, 135 (4), 501-512. [10.1037/0096-3445.135.4.501](#)
- Lin, M. H., Kwan, V. S. Y., Cheung, A. & Fiske, S. T. (2005). Stereotype content model explains prejudice for an envied outgroup: Scale of Anti-Asian American stereotypes. *Personality and Social Psychology Bulletin*, 31 (1), 34-47. [10.1177/0146167204271320](#)
- Liu, J., Harris, A. & Kanwisher, N. (2002). Stages of processing in face perception: An MEG study. *Nature Neuroscience*, 5 (9), 910-916. [10.1038/nn909](#)
- Macpherson, F. (2012). Cognitive penetration of colour experience. Rethinking the issue in light of an indirect mechanism. *Philosophy and Phenomenological Research*, 84 (1), 24-62. [10.1111/j.1933-1592.2010.00481.x](#)
- Macrae, C. N. & Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. *Annual Review of Psychology*, 51, 93-120.
[10.1146/annurev.psych.51.1.93](#)
- Macrae, C. N. & Martin, D. (2007). A boy primed Sue: Feature-based processing and person construal. *European Journal of Social Psychology*, 37 (5), 793-805.
[10.1002/ejsp.406](#)
- Macrae, C. N. & Quadflieg, S. (2010). Perceiving people. In S. Fiske, D. T. Gilbert & G. Lindzey (Eds.) *Handbook of social psychology* (pp. 428-463). New York, NY: McGraw-Hill.
- Markus, H. R. & Kitayama, S. (1991). Culture and the self. Implications for cognition, emotion, and motivation. *Psychological Review*, 98 (2), 224-253.
[10.1037/0033-295X.98.2.224](#)
- Martin, D. & Macrae, C. N. (2007). A face with a cue: Exploring the inevitability of person categorization. *European Journal of Social Psychology*, 37 (5), 37-5.
[10.1002/ejsp.445](#)
- Meeren, H. K. M., van Heijnsbergen, C. C. R. J. & de Gelder, B. (2005). Rapid perceptual integration of facial expression and emotional body language. *Proceedings of the National Academy of Sciences of the United States of America*, 102 (45), 16518-16523.
[10.1073/pnas.0507650102](#)
- Meltzoff, A. N. & Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science*, 198 (4312), 75-78. [10.1126/science.198.4312.75](#)
- (1994). Imitation, memory and the representation of persons. *Infant Behaviour and Development*, 17 (1), 83-99. [10.1016/0163-6383\(94\)90024-8](#)
- Moskowitz, D. S., Suh, E. J. & Desaulniers, J. (1994). Situational influences on gender differences in agency and communion. *Journal of Personality and Social Psychology*, 66 (4), 753-761.
[10.1037/0022-3514.66.4.753](#)
- Nagy, E. (2008). Innate intersubjectivity: Newborn's sensitivity to communication disturbance. *Developmental Psychology*, 44 (6), 1779-1784. [10.1037/a0012665](#)
- Nelson, C. A. (2001). The development and neural bases of face recognition. *Infant and Child Development*, 10

- (1-2), 3-18. [10.1002/icd.239](https://doi.org/10.1002/icd.239)
- Newen, A., Welpinghus, A. & Juckel, G. (forthcoming). *Emotion recognition as pattern recognition: the relevance of perception*.
- Newen, A. & Schlicht, T. (2009). Understanding other minds. A criticism of Goldman's simulation theory and an outline of the person model theory. *Grazer philosophische Studien*, 79 (1), 209-242.
- Newen, A. & Vogeley, K. (2003). Self-representation: Searching for a neural signature of self-consciousness. *Consciousness & Cognition*, 12 (4), 529-543. [10.1016/S1053-8100\(03\)00080-1](https://doi.org/10.1016/S1053-8100(03)00080-1)
- (2011). Den anderen verstehen. *Spektrum der Wissenschaft*, 8
- Norman, W. T. & Goldberg, L. R. (1966). Raters, ratees, and randomness in personality structure. *Journal of Personality and Social Psychology*, 4 (6), 681-691. [10.1037/h0024002](https://doi.org/10.1037/h0024002)
- Oyserman, D., Sakamoto, I. & Lauffer, A. (1998). Cultural accommodation: Hybridity and the framing of social obligation. *Journal of Personality Psychology*, 74 (6), 1606-1618. [10.1037/0022-3514.74.6.1606](https://doi.org/10.1037/0022-3514.74.6.1606)
- Panksepp, J. (2005). Affective consciousness: Core emotional feelings in animals and humans. *Consciousness and Cognition*, 14 (1), 30-80. [10.1016/j.concog.2004.10.004](https://doi.org/10.1016/j.concog.2004.10.004)
- Pegna, A. J., Khateb, A., Michel, C. M. & Landis, T. (2004). Visual recognition of faces, objects, and words using degraded stimuli: Where and when it occurs. *Human Brain Mapping*, 22 (4), 300-311. [10.1002/hbm.20039](https://doi.org/10.1002/hbm.20039)
- Phillips, W., Baron-Cohen, S. & Rutter, M. (1992). The role of eye contact in goal detection. Evidence from normal infants and children with autism or mental handicap. *Development and Psychopathology*, 4 (3), 375-383. [10.1017/S0954579400000845](https://doi.org/10.1017/S0954579400000845)
- Reddy, V. (2008). *How infants know minds*. Cambridge, MA: Harvard University Press.
- Roberts, T. & Bruce, V. (1988). Feature saliency in judging the sex and familiarity of faces. *Perception*, 17 (4), 829-840. [10.1068/p170475](https://doi.org/10.1068/p170475)
- Sander, D., Grandjean, D., Kaiser, S., Wehrle, T. & Scherer, K. R. (2006). Interaction effects of perceived gaze direction and dynamic facial expression: Evidence for appraisal theories of emotion. *European Journal of Cognitive Psychology*, 19 (3), 470-480. [10.1080/09541440600757426](https://doi.org/10.1080/09541440600757426)
- Schank, R. C. & Abelson, R. P. (1977). *Scripts, plans, oals and understanding. An inquiry into human knowledge structures*. New York, NY: Erlbaum.
- Schweinberger, S. R., Huddy, V. & Burton, A. M. (2004). N250r: A face-selective brain response to stimulus repetitions. *NeuroReport*, 15 (9), 1501-1505. [10.1097/01.wnr.0000131675.00319.42](https://doi.org/10.1097/01.wnr.0000131675.00319.42)
- Schyns, P. G., Bonnar, L. & Gosselin, F. (2002). Show me the features! Understanding recognition from the use of visual information. *Psychological Science*, 13 (5), 402-409. [10.1111/1467-9280.00472](https://doi.org/10.1111/1467-9280.00472)
- Secord, P. F., Dukes, W. F. & Bevan, W. (1954). Personalities in faces: I. An experiment in social perceiving. *Genetic Psychology Monographs*, 49 (2), 231-279.
- Stephens, G. L. & Graham, G. (2000). *When self-consciousness breaks: Alien voices and inserted thoughts*. Cambridge, MA: MIT Press.
- Sui, J. & Han, S. (2007). Self-construal priming modulates neural substrates of self-awareness. *Psychological Science*, 18 (10), 861-866. [10.1111/j.1467-9280.2007.01992.x](https://doi.org/10.1111/j.1467-9280.2007.01992.x)
- Tomasello, M. (1999). *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.
- Tranel, D. & Damasio, A. R. (1985). Knowledge without awareness: An autonomic index of facial recognition by prosopagnosics. *Science*, 228 (4706), 1453-1454. [10.1126/science.4012303](https://doi.org/10.1126/science.4012303)
- Uleman, J. S., Blader, S. L. & Todorov, A. (2005). Implicit impressions. In R. R. Hassin, J. S. Uleman & J. A. Bargh (Eds.) *The New Unconscious* (pp. 362-392). New York, NY: Oxford University Press.
- van den Stock, J., Righart, R. & de Gelder, B. (2007). Body expressions influence recognition of emotions in the face and voice. *Emotion*, 7 (3), 487-494. [10.1037/1528-3542.7.3.487](https://doi.org/10.1037/1528-3542.7.3.487)
- Vecera, S. P. & Johnson, M. H. (1995). Gaze detection and the cortical processing of faces: Evidence from infants and adults. *Visual Cognition*, 2 (1), 59-87. [10.1080/13506289508401722](https://doi.org/10.1080/13506289508401722)
- Vetter, P. & Newen, A. (2014). Varieties of cognitive penetration in visual perception. *Consciousness & Cognition*, 27, 62-75. [10.1111/j.1468-0017.2006.00298.x](https://doi.org/10.1111/j.1468-0017.2006.00298.x)
- Vogeley, K. (2012). *Anders sein. Hochfunktionaler Autismus im Erwachsenenalter*. Weinheim, GER: Beltz.
- Vogeley, K. & Newen, A. (2002). Mirror neurons and the self construct. In M. Stamenov & V. Gallese (Eds.) *Mirror neurons and the evolution of brain and language* (pp. 135-150). Amsterdam, NL: Benjamins.
- Vogeley, K. & Roepstorff, A. (2009). Contextualising culture and social cognition. *Trends in Cognitive Sciences*, 13, 511-516.

- Volz, K. G. (2008). Ene mene mu insider und outsider. In R. Schubotz (Ed.) *Other minds. Die Gedanken und Gefühle anderer* (pp. 19-30). Paderborn, GER: Mentis.
- Vosgerau, G. & Newen, A. (2007). Thoughts, motor actions and the self. *Mind and Language*, 22 (1), 22-43. [10.1111/j.1468-0017.2006.00298.x](https://doi.org/10.1111/j.1468-0017.2006.00298.x).
- Vuilleumier, P. & Pourtois, G. (2007). Distributed and interactive brain mechanisms during emotion face perception: Evidence from functional neuroimaging. *Neuropsychologia*, 45 (1), 174-194. [10.1016/j.neuropsychologia.2006.06.003](https://doi.org/10.1016/j.neuropsychologia.2006.06.003)
- Wicker, B., Keysers, C., Plailly, J., Royet, J.-P., Gallese, V. & Rizzolatti, G. (2003). Both of us disgusted in my insula: the common neural basis of seeing and feeling disgust. *Neuron*, 40 (3), 655-664. [10.1016/S0896-6273\(03\)00679-2](https://doi.org/10.1016/S0896-6273(03)00679-2).
- Williams, L. M., Palmer, D., Liddell, B. J., Song, L. & Gordon, E. (2006). The “when” and “where” of perceiving signals of threat versus non-threat. *NeuroImage*, 31 (1), 458-467. [10.1016/j.neuroimage.2005.12.009](https://doi.org/10.1016/j.neuroimage.2005.12.009)
- Willis, J. & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science*, 17 (7), 592-598. [10.1111/j.1467-9280.2006.01750.x](https://doi.org/10.1111/j.1467-9280.2006.01750.x)
- Wimmer, H. & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children’s understanding of deception. *Cognition*, 13 (1), 103-128. [10.1016/0010-0277\(83\)90004-5](https://doi.org/10.1016/0010-0277(83)90004-5)
- Yamaguchi, S., Kuhlman, D. M. & Sugimori, S. (1995). Personality correlates of allocentric tendencies in individualist and collectivist cultures. *Journal of Cross-Cultural Psychology*, 26 (6), 658-672. [10.1177/002202219502600609](https://doi.org/10.1177/002202219502600609).

Multiplicity Needs Coherence – Towards a Unifying Framework for Social Understanding

A Commentary on Albert Newen

Lisa Quadt

In this commentary, I focus on Albert Newen’s multiplicity view (MV) and aim to provide an alternative framework in which it can be embedded. Newen claims that social understanding draws on at least four different epistemic mechanisms, thus rejecting the idea that there is a default mechanism for social cognition. I claim that MV runs the risk of combining elements that have been described in meta-physically incompatible theories. I will argue that multiplicity needs coherence, which can be achieved by applying the theoretical framework of first-, second-, and third-order embodiment (1-3E; [Metzinger 2014](#)) to the study of social cognition. The modified version of this theory, 1-3sE (first-, second-, and third-order *social* embodiment), can serve as a unifying framework for a pluralistic account of social understanding.

Keywords

Direct perception | Embodiment | Interaction | Interactive turn | Mirror neurons | Multiplicity view | Phenomenology | Social cognition | Social understanding

Commentator

[Lisa Quadt](#)

lisquadt@students.uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Target Author

[Albert Newen](#)

albert.newen@rub.de
Ruhr-Universität Bochum
Bochum, Germany

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Introduction

The multiplicity view (MV) is part of Newen’s person model theory (PMT) and claims that individuals apply multiple epistemic strategies to make sense of other people, namely simulation, theoretical inference, direct perception (DP) and primary interaction.¹ He thus interestingly argues against the view that there is

something like a default strategy of social understanding. In the following, I will scrutinize MV and, in doing so, attempt to reach three goals: First, I reconstruct the main claims of MV and suggest that the development of such a pluralistic account of social cognition can be seen as contributing to the so-called “interactive turn” ([Overgaard & Michael 2013](#); section “The multiplicity view”). MV has the potential

¹ For a brief explanation of the terms, see [Newen this collection](#), pp. 1-2.

to integrate bodily and interactive contexts, while also paying more attention to the phenomenology of social encounters. Second, I argue that current pluralistic depictions of social cognition – of which MV is a clear example – run the risk of operating under (often implicit) contradictory background assumptions. In the section “Multiplicity needs coherence”, I first show how and why different social cognitive mechanisms have been described under different sets of metaphysical assumptions. Since these assumptions are often contradictory, a coherent version of MV cannot simply claim to combine them. I then go on to argue that the concept of DP as an epistemic mechanism is either metaphysically incompatible with simulation and theorizing, empirically implausible, or – if it is re-formulated so that it fits a representationalist description – does not meet the goal of integrating embodiment and phenomenology anymore. I will thus claim that DP should be used as a phenomenological rather than epistemological concept. My third goal is then to suggest novel ways of adopting a pluralistic perspective on social cognition, while remaining in metaphysically coherent territory. Metzinger’s theory of first-, second-, and third-order embodiment (1-3E) is a conceptual framework that combines representationalist and non-representationalist levels of analysis in order to show how a specific phenomenal quality (e.g., phenomenal selfhood) can arise within an embodied system (Metzinger 2014). Metzinger claims that phenomenal properties are computationally grounded in a representation of one’s body (the “body model”, *ibid.*, p. 273), which in turn is physically implemented by bodily and neural structures. I aim to apply this idea to the study of social understanding (section “1-3sE – Levels of social embodiment”). This application enables a more fine-grained depiction of different phenomenal qualities in social encounters and shows their putative relation to representational and physical counterparts. I ask which parts of the body model could potentially be shared and thus be exploited for a skillful navigation of an individual’s social environment. In a last step, I sketch the physical grounds of social cognition.

2 The multiplicity view

The multiplicity view (MV) is part of Albert Newen’s person model theory (PMT), which provides a rich and detailed account of social understanding. It attempts to answer two central questions in the research field of social cognition, which the author neatly differentiates and then again integrates into a comprehensive theory. The first question asks which epistemic strategy humans use to access the mental states of others and to gather information about them. Approaches advocating Simulation Theory (ST; e.g., Goldman 2006), as well as direct perception (DP; e.g., Gallagher 2008), have attempted to yield an answer to that question, while Theory Theory (TT; e.g., Gopnik & Meltzoff 1997) and Narrative Practice Hypothesis (NPH; e.g., Hutto 2008) focused on a second question: How is the information we obtain to understand others stored and organized? By sorting out these questions, Newen shows that different theories have tried to tackle different problems, which I believe to be a very useful and fruitful contribution to the research field. It reveals that the four main theories mentioned above are less competitive than originally thought, since, on closer examination, they actually aim to give answers to different questions. This viewpoint enables one of Newen’s main arguments, namely that each of these approaches can be merged into one unified account of social understanding. He takes three steps in arguing for his theory. In a first step, he differentiates between the two questions in the research field of social cognition mentioned above, thus setting up a dividing line between the vast manifold of different approaches and theories. Secondly, the author puts forth a pluralistic account of social cognition, the multiplicity view (MV). In doing so, he attempts to answer the first question discussed earlier. In a third step, Newen tackles the second question of how knowledge about other people is organized and stored. He claims that this happens through the formation of so-called *person models*, hence person model theory (PMT; see Newen this collection).

By laying out MV as a pluralistic account of social cognition, Newen aims to steer the discussion in the research field into a different direction, away from debating whether social understanding is a form of simulation, theoretical inference, DP or interaction. Instead, he argues that all four epistemic strategies are applied, depending on the social context (cf. [Newen this collection](#), p. 7). MV is of particular interest, because it reflects two growing convictions in the research field. First, by paying attention to DP and interaction, it does justice to demands that arose in the so-called “interactive turn” ([Gallotti & Frith 2013](#); [Overgaard & Michael 2013](#)) and can thus be seen as part of the movement itself. The interactive turn claims that researchers have not paid enough attention to the phenomenology of social encounters ([Gallagher 2001](#)), the interactive contexts in which most social situations are embedded ([De Jaegher & Paolo 2007](#)) and the role of the body and emotions in social cognition ([Schilbach et al. 2013](#)). This directly relates to MV, since it aims to include intuitive ways of social understanding that do not necessarily require simulation and theoretical inference and thus to widen the theoretical scope towards less “cognitivist” views. The second conviction is that there is more to social cognition than a single all-purpose mechanism ([Adolphs 2006](#), p. 30; [Fiebig & Coltheart in press](#)).² (Human) social cognition obviously is manifold; it has many aspects that are not only phenomenologically distinct (just think of the different experiences you have when trying to figure out your advisor’s somewhat cryptic Email, or when trying to make your 4 year-old eat her spinach), but also draws on several cognitive mechanisms that are differently implemented. It therefore makes sense that we can find something useful in each of the four theoretical approaches discussed so far; while ST and TT are plausible accounts to describe and explain “higher-level” social cognition that requires

quite sophisticated skills, other theories such as DP or interaction theory cover more intuitive ways of understanding others. Merging them into a comprehensive theory seems to be a natural next step.

[Newen](#) claims that

[t]here is no standard default strategy of understanding others, but in everyday cases of understanding others we rely on a multiplicity of strategies which we vary depending on the context and on our prior experiences (and eventually also triggered by explicit training). ([this collection](#), p. 7)

How does he arrive at this conclusion? Newen argues against the view that only *one* of the mechanisms that have been proposed to be important for social cognition (simulation, theorizing, DP and primary interaction) can plausibly be viewed as the default strategy by which humans understand each other. The main argument against such a single-mechanism view is that their activation seems to be highly context-dependent. Simulation, according to Newen, presupposes similarity between two interacting individuals. Theorizing only applies in complex social situations which need explicit and thoughtful disambiguation. Encountering someone of whom we already have rich prior information activates DP, while social situations that are easy to understand can be disambiguated by primary interaction. Thus, [Newen](#) concludes that “[o]nly the combination of all four strategies, in full sensitivity to the context and applied on the basis of our experience in successfully using the strategies, makes us experts in understanding others” (*ibid.*, p. 7).

3 Multiplicity needs coherence

While this surely is an attractive way to describe social understanding, and does justice to its oft-proclaimed manifoldness, these mechanisms have been described in several theoretical frameworks that operate under different (and partly contradictory) metaphysical

² Such a view can already be found in Goldman’s work. He endorses a hybrid account of mindreading, which describes “a number of ways to blend simulation and theorizing elements into a mosaic of mindreading possibilities” ([Goldman 2006](#), p. 43).

background assumptions.³ Thus, a simple combination of them does not come easily. Simulation and theory-based inference have been described within a computationalist, cognitivist framework which often assumes that the mind is mainly a representational and internal device (Bruin & Kästner 2012), i.e., a functional structure locally realized in the brains of individual organisms. Bodily and environmental structures play at most an enabling or causal role for a specific *internal* mechanism. In contrast, DP and primary interaction, both of which are concepts stemming from the phenomenological tradition, have their roots in an enactive account of cognition (cf. Gallagher 2008, p. 537), thus rejecting basic metaphysical assumptions of cognitivism (e.g., representationalism, reductionism, mechanistic explanations; Rowlands 2009).⁴ The theoretical background of DP and primary interaction views the mind as a non-representational, relational device which emerges within the skillful interaction between organism and environment:

The enactive interpretation is not simply a reinterpretation of what happens extraneurally, out in the intersubjective world of action where we anticipate and respond to social affordances. More than this, it suggests a different way of conceiving brain function, specifically in nonrepresentational, integrative and dynamical terms. (Gallagher et al. 2013, p. 422)

More specifically, enactive and phenomenological approaches to social cognition not only see the body as part of cognitive processing, they also assign a very important status to interaction. While enactive theories display interaction as (at least possibly) *constituting* social cogni-

ive processes (De Jaegher & Paolo 2007, p. 493), traditional mindreading theories have not even considered interaction to be an element which could influence social cognition (cf. Fuchs & Jaegher 2009, p. 466).

There are several reasons why ST and TT have been spelled out in a more cognitivist set of assumptions, while DP and primary interaction have been described in reference to an enactive framework. Although their roots in the history of ideas plays an important role, there are deeper systematic reasons why it makes sense to couch them in different sets of metaphysical assumptions. To see this, consider the relation between the external world and internal processing in either framework. A rather cognitivist view assumes that the task of the brain is to figure out the outside world and that this world is *internally represented*.⁵ Since other people belong to this world outside of one's own mind, it follows that the causes for their behavior need to be inferred by internal representation processing as well. Because it is assumed that the brain is the only mental organ (Hohwy [this collection](#)), the *location* of (social) cognitive processing thus can be said to be inside one individual's head. Simulation and theorizing fit neatly into this picture of the mind; they are inference processes which function to disambiguate social input and are implemented by specific neural mechanisms. By contrast, an enactive view of social cognition as has been described by De Jaegher and colleagues and advocated by Gallagher, presupposes two different assumptions. First, in order to assume that interaction dynamics carry as much of the "cognitive load" to understand other minds as is proposed, a relational view of the mind enters the picture. It is important to understand that an enactive view is not the same as an externalist view, which could be compatible with assumptions of

³ I am well aware of the fact that there are many shades of both cognitivist and enactive views. I will therefore focus on the views of the authors that have been cited by Newen in the target paper. For a general introduction, see for example Thompson (2010); Varela et al. (1993); Rowlands (2009).

⁴ The difference between enactive and phenomenological theories seems to boil down to the explanatory scope. While enactivism explicitly claims to offer a radically different alternative to cognitivism and thus builds a proper account of cognition (Varela et al. 1993) phenomenology is mostly seen as a description of experiential phenomena (Gallagher 2008).

⁵ Although this seems to be a rather "old" view, it is currently celebrating a comeback. Jakob Hohwy, for example, claims that the consequences of advocating predictive processing (2013; see also Clark 2013a) are to adopt a fully internalist picture of the mind. In his words, there is an "evidentiary boundary" (Hohwy 2014, p. 6) between what has to be inferred (viz., hidden causes in the external world) and the inference device (the brain). Accordingly, all the processing takes place within this boundary, which happens to be the skull (cf. *ibid.*, p. 8). Please note, though, that both Clark and Seth propose a more embodied perspective on prediction (Clark [this collection](#); Seth [this collection](#)).

the cognitivist camp (cf. Rowlands 2009, p. 54). The mind is, according to such an enactive perspective, neither internal nor external; it constitutes itself within the relation (hence *relational*) between an embodied agent and its environment (cf. Di Paolo & Thompson 2014, p. 68; Engel et al. 2013, p. 202). Such a view enables the claim that interactions are examples of this unfolding mental process and thus constitute social cognition. This claim is incompatible with an internalist perspective, which does not ascribe any constitutional power to mind-external properties.

Furthermore, if the external world and the minds of others could be *directly* perceived without further mental processing or inference, neither simulation nor theoretical inference would be needed. This is exactly the point of the non-cognitivist camp, as becomes obvious in this quote by Newen: “The mental states of others are not hidden, and need not to be inferred on the basis of perceiving the behavior; rather, behavior is an expression of the mental phenomena that, in seeing the behavior, is also directly seen” (this collection, p. 5). What does it mean that something can be *directly* seen? Gibson (1979) introduced DP in relation to his famous conception of “affordances”: “The affordances of things for an observer are specified in stimulus information. They seem to be perceived directly because they are perceived directly” (Gibson 1977, p. 79). Importantly, the direct perception of affordances is possible because, according to Gibson, affordances are physically *real* (i.e., they exist independent of the perceiving subject) and as such are perceivable properties of objects in the environment (cf. 1979, p. 129). Note how this is crucially different from a view which assumes that object properties need to be mentally represented, thus requiring an intermediary step.⁶ However, Gallagher makes explicit in a footnote (cf. 2008, p. 537) that his conception of DP is not to be *entirely* equated with a Gibsonian notion of the term. Gallagher emphasizes that he does not deny the underlying

complexity of perceptual processing, much rather he counts those processes as belonging to perception. He thus puts forth the conception of “smart perception”:

But this informing process is already built into the perceptual process so that as I consciously perceive, my perception is already informed by the relevant sub-personal processing. I don’t first perceive and then add memory in order to recognize my car. My perception, in this sense, is direct even if the sub-personal sensory processing that underpins it follows a complex and dynamic route. (*ibid.*, p. 537)

Even with that kind of definition, his view still presupposes that there are properties of external objects that can be “directly” picked up, that exist independently from the perceiving subject. As such, it is indeed *reminiscent* of a Gibsonian conception. The difference between cognitivist and non-cognitivist pictures of social cognition, in the cases that I just described, seems to boil down to the metaphysical assumption of whether or not there are hidden causes in the outside world that require an inference or representational mechanism in order to access and process them. While ST and TT clearly assume such a view, DP denies it. Therefore, I claim that MV cannot simply combine theoretical elements that draw on such considerable metaphysical differences.

Another important difference between these theoretical approaches is how each treats the issue of phenomenology. While the experiential nature of social encounters plays at most a minor role in mindreading theories, such as ST and TT, the phenomenal level is of paramount importance for the enactive camp, who advocate for DP. This becomes most obvious in the claim that the experienced smoothness and immediacy of social interactions tells us something about the epistemic access to other minds. However, “directness” as a concept in academic research is relative to a specific level of description. Let me explain this in more detail. Consider Gallagher’s argument that smart perception is a subpersonally informed mechanism (cf.

⁶ In the following, I will use the requirement of intermediary steps as the distinctive feature that differentiates directness and indirectness. In doing so, I follow De Vignemont: “There is a direct access if and only if the causal transmission of information is direct and does not involve intermediary steps” (2010, p. 291).

2008, p. 537) that directly enables an individual to perceive the minds of others without “additional mental effort.” It is based largely on the rapid activation of mirror neurons (30-100ms, *ibid.*, p. 541), such that he claims a distinction between a merely perceptual process and an additional mental process does not make sense. In his words:

A distinction at the neural level between activation of the visual cortex and activation of the pre-motor cortex does not mean that this constitutes a distinction between processes that are purely perceptual and processes that involve something more than perception. (*ibid.*, p. 541)

The question that follows is how one should individuate mental mechanisms, and I suggest that *functional properties* are much more substantial and conceptually relevant individuation criteria than temporal properties. It is, to me, highly questionable whether temporal correlation justifies assuming that there is mechanistic inseparability. The functional role of a mental mechanism seems a much less arbitrary criterion. Furthermore, it enables a more fine-grained view of the subpersonal processes that underlie social cognition. Instead of talking about perception—which could include all processes if only they are activated in a more or less specific amount of time—it is possible to take a closer look at which brain region correlates with which mechanism. If mechanisms are individuated by their *functional role* instead of the temporal properties of the physical realizers of this functional role, it makes sense to assume that the visual system and the mirror neuron system are distinct. If they are, however, it is unfeasible to speak of “smart perception”. This concept presupposes that perceptual and post-perceptual processes can coherently be described as *one* mechanism, which I reject. Additionally, the concept of “direct perception” does not apply anymore either, since mirror mechanisms should be seen as a functionally distinct and therefore intermediary step in the process of understanding others. I thus conclude that DP—as described by Gallagher—does not co-

herently apply to the subpersonal level of description.

This relates to my main point, namely that there are different levels of description at which a phenomenon can be scrutinized. At the phenomenological level, DP can be described as the experience of *directly* and *immediately* perceiving the other person’s mental states. I walk into my living room, I see my friend’s face and I experience myself as instantaneously knowing that she is really upset. However, this experiential quality of directness is brought forth by a subpersonal process, which is indirect, as I have argued above. At any other level of description, therefore, directness does not apply. In this view, DP is a phenomenal quality of some mental states and should thus not be confused with the epistemic mechanism *itself*. The simultaneity in our everyday experience does not justify anything on other levels of description. I therefore argue that DP should be treated as a phenomenal quality of *some* social encounters instead of assigning it the status of an epistemic strategy to access other minds.

Note that Newen does not explicitly support a phenomenological or enactive view of the mind, nor does he make any claims about the metaphysics of social cognition. What he does do, however, is emphasize Gallagher’s conception of DP and primary interaction as being the main sources for an epistemic access to other minds (cf. Newen *this collection*, p. 8). If Newen was to reject the strong claims of a non-representational view of (social) cognition, however, it is questionable how closely his notions of DP and primary interaction, as core concepts of his theory, actually relate to their original formulations. This leaves us with two options. The first is to assume that Newen fully endorses the views of his oft-cited colleague. In this case, the problem of compatibility becomes obvious. The second, and more likely possibility is that the author does not support DP and primary interaction with all their metaphysical implications. It indeed seems that he rather re-formulates both concepts so that they possibly fit into a representational framework. According to Newen (*this collection*, p.5), DP is realized by a process of pattern recognition and primary in-

teraction – although Newen explicitly cites [Gallagher & Hutto \(2008\)](#) – is characterized as follows: “[...] I notice a social act being directed towards me and so start to interact, such that a standard interaction is realized, which may be nonlinguistic but may also involve linguistic communication [...]” ([Newen this collection](#), p. 7). What is problematic here is that one of the most interesting and valuable features of MV gets lost, namely its potential to fulfill demands of the interactive turn. A true fulfillment would require widening the theoretical scope of social cognition by going beyond the study of individual brains and considering bodily, interactive and phenomenological processes more carefully.

What needs to be reconciled and made conceptually consistent is thus our choice of a specific, unified methodological framework—our overarching theoretical approach of simulation, theory-based inference, DP and primary interaction—since they all describe important aspects of social understanding. It should be a common aim to work with a coherent set of metaphysical assumptions, since whether or not one agrees on either set of background assumptions has important implications for both theoretical and empirical research. Not only does that decision influence our choice of the *unit of analysis*, i.e., how we frame the explanatory unit for empirical research. For a long time, this unit has been one individual observing another. It has been claimed, however, that this does not properly reflect the real nature of social cognition, and thus a shift is needed:

The explanatory unit of social interaction is not the brain, or even two (or more) brains, but a dynamic relation between organisms, which include brains, but also their own structural features that enable specific perception-action loops involving social and physical environments, which in turn affect statistical regularities that shape the structure of the nervous system [...]. ([Gallagher et al. 2013](#), p. 422)

When an enactive or phenomenological perspective is adopted and the status of interaction as constituting social cognition is accepted, this

adds an additional *level of analysis* (i.e., an “interactionist stance”; [De Jaegher et al. 2010](#)) while erasing one that is profound and fundamental for most researchers: representation. Furthermore, the shared goal to pay more attention to the body, interaction and phenomenology comes with many methodological challenges. For all these reasons it should be in the common interest of the research field to find a way to ease the tensions.

As I have shown, Newen tries to combine four elements that might not be entirely compatible. However, the core of his idea is highly valuable, and certainly should not be rejected. What his pluralistic account of social cognition claims is that there are low-level social mechanisms that mainly rely on interaction and do not need complex or explicit thought, while higher-level, sophisticated mechanisms play a just as important role for the phenomenon. While some social situations require processes that allow complex thinking, other contexts can be intuitively disambiguated. In what follows, I will sketch an alternative framework, based on Metzinger’s theory of three-level embodiment, which I claim is able to integrate the four elements while operating on coherent background assumptions. Additionally, it has the potential to fulfill the demands of the interactive turn by paying more attention to interactive contexts, the role of the body and the importance of phenomenology.

4 1-3E – First-order embodiment, second-order embodiment, third-order embodiment

Before I describe how the framework of 1-3E itself can be exploited for a pluralistic picture of social cognition, let me describe the framework in more detail. Metzinger’s goal is to provide a framework which shows how the experience of being a self is generated within an embodied system (cf. [Metzinger 2014](#), p. 272). The basic assumption is that experiential phenomena (such as phenomenal selfhood) can be described at several different levels: they have a specific phenomenal quality (i.e., phenomenological level of description), which is brought forth by under-

lying computations and representations (i.e., computational/representational level of description). These are implemented by their physical counterparts (i.e., implementational level of description). 1-3E is a theory about the grounding relations between them, that is, the grounding relations holding between phenomenal properties of representational states and their physical and computational resources. In a broader context, Metzinger claims that “the self” is not a thing or an entity (2004), but rather the phenomenal product of a complex computational process which happens to take place in embodied systems. If that is the case, however, the following question arises: How exactly is the experience of being a self generated within an embodied system? In other words, what are the grounding relations of phenomenal selfhood?⁷

Metzinger introduces three levels: first-order embodiment, second-order embodiment and third-order embodiment (Metzinger 2006, 2014). Importantly, these concepts not only describe different levels of embodiment and their relation *within* one system, they also refer to different *classes of systems* which possess different kinds of embodiment. To see this, think of the following three systems which all possess a body and some sort of skillful behavior: a worm, an advanced robot (e.g., the “starfish”, see Metzinger 2007), and a human in a waking state. As for the worm, it is safe to say that, in order to navigate its environment, it directly exploits its physical (i.e., bodily) resources. It is highly unlikely, however, that one would find any rule-based computation over an explicit symbol-like representational structure in the worm’s nervous system. In Metzinger’s terms, this kind of system possesses first-order embodiment (1E system). In contrast to this rather rudimentary kind of embodiment, 2E systems (i.e., systems which possess second-order embodiment) do unconsciously represent themselves *as* embodied. This means that they have some kind of body model that can be exploited by the system in

several ways (e.g., as a functional tool for motor control) and sustains skillful interaction with the environment. Importantly, 2E enables counterfactual representation, i.e., the ability to represent possible states without actual execution. The body model thus functionally underlies both physical and virtual behavior (see Cruse & Schilling this collection). What 2E systems are lacking, however, is a *phenomenal* representation of themselves *as* embodied systems. While a robot like the starfish can *use* its unconscious body representation to steer movements, it does not *experience* itself as doing so. Only systems that possess third-order embodiment (3E systems) experience this phenomenal quality of being an agent that owns a body. Humans in non-pathological waking states, for example, possess this kind of embodiment. Along with the ability to use their body model in the same way as 2E systems do, they have the additional sense of owning and controlling this model (cf. Metzinger 2014, pp. 274–275). Interestingly, it is also here that we once again find the phenomenology of “directness” and “immediacy”. It is important to note that 2E and 3E systems always possess lower levels of embodiment as well, since they build onto each other and higher levels presuppose the existence of lower levels. In this way, 1-3E can be seen as a grounding theory. To briefly summarize, systems that phenomenally represent themselves as embodied agents possess 3E. Phenomenal properties of states, described at this level, are computationally grounded by referring to a unified representation of the body – second-order embodiment. This unconscious body model, in turn, is grounded in physical and bodily resources, which are described at the lowest level of the hierarchy.⁸

Metzinger is clear about the relation between 2E and 3E; the representational content

⁷ “It is the problem of describing the abstract computational principles as well as the implementational mechanics by which a system’s phenomenal self-model (PSM; cf. Metzinger 2003, 2007) is anchored in low-level physical dynamics, in a maximally parsimonious way, and without assuming a single, central module for global self-representation.” (Metzinger 2014, p. 272)

⁸ I have argued before that a simple combination of cognitivist, representational, and enactive, non-representational perspectives results in a metaphysically incoherent view. One could ask why it should now be possible for 1-3E to put together non-representational and representational levels of description. As I have described earlier, most enactive theories reject representations entirely (e.g., Fuchs & Jaegher 2009, p. 466). That is one important reason why such a view is incompatible with representational theories. Grounding theories, however, take a different perspective on representations. They view them as *grounded* in bodily processes (cf. Pezzulo et al. 2013, pp. 6). As such, representations can be seen as a phenomenon that gradually emerges within an embodied system (cf. Metzinger 2014, p. 278).

of 2E is “elevated to the level of global availability and integrated with a single spatial situation model plus a virtual window of presence” (2014, p. 274). However, one thing that remains relatively vague in his theory is the relation between 1E and 2E. The problem I see here is that Metzinger does not explicitly describe what actually grounds 2E and which role bodily structures play besides that of yielding a grounding relation.⁹ A 1E system is defined as a “purely physical, reactive system”, which adapts to its environment by exploiting its physical resources. This is not, in my view, what is being represented by a 2E system, which represents itself “as an embodied agent” (*ibid.*, p. 273). What is needed is a more detailed and specific description of 1E and its relation to 2E. Therefore, the discussion of 1E in my own proposal is twofold. First, I analyze the low-level mechanisms that can be described at this level, claiming that they enable basic social skills (e.g., coupling). Second, I describe which neural, bodily and perhaps even extra-bodily structures most likely underlie social processes that are located at the level of 2sE.

There is one important aspect of 3E that I wish to describe in more detail as it will be crucial for my theory. Metzinger distinguishes two kinds of phenomenal properties instantiated by conscious representational states; they can be either *transparent* or *opaque*. Notice that he uses those terms in a rather counterintuitive way I will try to make sense of in the following.¹⁰ An analogy that might help to do so is to think of the difference between a freshly cleaned and a quite dirty window front. In the first case, when the glass is transparent, we can see everything behind it while not perceiving the glass *as* a medium we are looking through. However, if the glass is dirty and opaque, we will not only have trouble seeing the things behind it, we will also perceive the window *itself as* something we are looking through.¹¹ In analogy, consider mental states (and

their processing stages) as either transparent or opaque. A mental state is opaque when it is experienced *as* a representational state. A quite straightforward example is explicit thought where an individual is consciously aware of the fact that she is thinking. The process of representation *is represented as such* in this case, and is therefore opaque. In contrast, if a state is transparent, earlier processing stages are not phenomenally represented; they are not part of the experience of an individual. In the case of phenomenal selfhood, for example, all that is experienced is the sense of being a self in a world. The fact that this experience is a representational process is not part of its phenomenal content. Note that the distinction between phenomenal properties of epistemic mechanisms (such as computations and representations) and epistemic mechanisms themselves is central to the concept of transparency. If we do not experience that a specific phenomenal state is generated subpersonally, when the underlying processes are not elevated to the level of experience, all we experience is the subjective, phenomenological profile of that state. Such a claim is only valid, however, if we assume that these two levels are actually distinct, which seems to be denied by some philosophers in the phenomenological tradition.

In what follows, I will modify parts of the 1-3E framework in order to make it suitable for a pluralistic view of social understanding. The basic scaffold of the theory is retained, since its hierarchical structure is helpful for describing a multi-faceted phenomenon like social cognition. It also offers the possibility for future research to pair 1-3E and 1-3sE with other hierarchical theories of cognition, such as the predictive processing framework (PP; Clark 2013b; Hohwy 2013). PP has not only been described as a very promising theory to unify perception, action and cognition (Clark 2013b), it has also been fruitfully applied to social cognition (Kilner et al. 2007). 1-3sE has the potential to integrate this explanatorily powerful approach, the details of which can be spelled out in future research, but cannot be pursued in this commentary. I furthermore adopt the idea that different levels

don’t see the window, but only the bird flying by.” (2003, p. 358)

⁹ He gives, though, an example of phenomenal dream states, showing how (parts of) the body model is grounded in bodily structures and processes. Physical eye movements, in this case, most likely ground the phenomenal experience in lucid dreaming (cf. Metzinger 2014, p. 276).

¹⁰ For a more detailed description of former usage of the terms, see Metzinger 2003, pp. 345–358.

¹¹ Metzinger uses a similar example: “With regard to the phenomenology of visual experience transparency means that we are not able to see something, because it is transparent. We

of embodiment represent different levels of sophistication and complexity in a system. In order to strengthen this idea and to give an even more differentiated view of social understanding, I aim to make the difference between transparent and opaque social states more obvious. While the general distinction between transparency and opacity is retained, I will modify this aspect in order to make it fruitful for social understanding. To do so, I introduce the concept of “3sE+”, which describes experiences in social situations that need explicit and conscious thinking.

Transparency makes it furthermore possible, according to Metzinger, to distinguish one’s own body from that of others (cf. Metzinger 2014, p. 274). However, there is an objection I wish to make about this point. I claim that a self-other distinction that functionally serves to identify one’s own body in contrast to those of others is already present at the level of 2sE and thus can be achieved without *phenomenally* representing one’s body. I will argue for this claim in more detail in the next section.

Additionally, my proposal offers novel ways to enrich Metzinger’s original account. He claims that the functional structure of the body model opens a window into social cognition (cf. *ibid.*, p. 273). However, I suggest that this could be a bidirectional relation. There are hints in the literature that being immersed in a social environment is crucial and formative for more general cognitive skills and their development. For example, anecdotal evidence shows that emotional neglect of caregivers severely impairs the physical and mental development of children (Zimmer 1989). Empirical research furthermore shows that the presence, interaction, perception and emotional engagement of and with others shape self-related body representations (e.g., Furlanetto et al. 2013; Schilbach et al. 2013). Longo & Tsakiris (2013) thus conclude that this line of research suggests a strong connection between first-person and so-called second-person (Schilbach et al. 2013) processes, which needs to be considered by researchers of each camp: “Such findings support a model of first-person perspective according to which our sense of self is plastically affected by multisensory informa-

tion as it becomes available during self-other interactions” (Longo & Tsakiris 2013, p. 430). I thus conclude that it should not only be considered how the development of a self-model influences social cognition, but also which role social processes play in forming such a self-model. This opens interesting and new questions for research on both social cognition and the self. One could ask, for example, whether some social cognitive skills are necessary for the development of a stable self-model or whether there are “genuinely social” parts of the self-model.

5 1-3sE– Levels of social embodiment

In this section, I will introduce an alternative framework in which I describe different processing stages of social understanding as different levels of social embodiment. Before I go into detail about how to apply 1-3E to social understanding, let me motivate my strategy here. I have already pointed out why MV yields an attractive theoretical assumption for research on social cognition. It allows, to briefly repeat, the integration of different aspects of a manifold phenomenon and thus aims to give a comprehensive perspective that is able to encompass sub-areas of interest and research. The advantage of couching MV into 1-3E is that its hierarchical nature affords this integration at different levels of description, while operating on a set of coherent background assumptions. As a grounding theory, it suggests how different levels of analysis relate and at the least has the potential to assign an important role to aspects that lay outside an individual brain. As such it can also do justice to demands from the interactive turn, viz. the consideration of interaction dynamics and their possible role for social cognition as well as taking the phenomenology of social encounters seriously. However, MV suffers from the problem of metaphysical incompatibility. 1-3E, on the other hand, is a representational account that offers a metaphysically sound ground for a manifold phenomenon. My goal is to scaffold a framework for human social cognition, which, as I will argue, can be described as a case of 3E in non-pathological human individuals.

I will now briefly give a rough overview of my proposal of a three-level model of social understanding which I dub “1-3sE” (first-order social embodiment, second-order social embodiment, third-order social embodiment)¹², before I go into detail about what each level amounts to. As in the original version of the framework, levels of social embodiment represent both levels *within* a system and different *kinds* of systems. I thus assume that each social third-order system possesses first- and second-order social embodiment, too. In this commentary, I will focus on describing levels of embodiment within social systems, since this aspect of the framework is of greater importance for a pluralistic view of social cognition.

As previously mentioned, I take it that 1sE fulfills a twofold function: First, it serves as the implementational level of description, showing which physical parts ground higher-level, representational and phenomenal processes. Second, low-level sensorimotor mechanisms subserve basic social interactions (e.g., coupling or synchronization). 2sE involves the instantiation of a model which pre-reflexively represents features of the body. It is assumed that parts of this body model can be shared and thus functionally underlie social cognitive processes that may well operate at the unconscious level, such as imitation, joint attention and action understanding. Finally, 3sE describes cases of consciously experienced social understanding. I claim that there are various kinds of phenomenal experiences in social situations that can be differentiated by applying the concepts of transparency and opacity. Since I consider opaque social mental states to exhibit a very special kind of experience, which is not only rare, but might also entail an additional level of representation, I introduce an extra level: 3sE+. I will now describe the specific levels and their relation in more detail, before I show how my view overcomes the shortcomings of MV.

5.1 Third-order social embodiment (3sE)

Individuals that phenomenally represent themselves as social individuals can be described as

social 3E systems (3sE). There are certainly many different ways in which humans experience themselves as being social, but I will focus on those that are mentioned by Newen: DP, personal-level simulation, and explicit theoretical inference.

The concepts of transparency and opacity allow a more fine-grained distinction of different phenomenal experiences of social encounters, as they offer a way to emphasize the similarities and differences between various phenomenal qualities in social situations. DP describes the experience that I can, without being aware of any intermediary steps, understand another person. Importantly, as Zahavi points out, the perceived directness still holds in cases of “unsuccessful” social understanding, such as deception or misunderstandings (cf. 2011, pp. 548–549). Although I can get what you say completely wrong, for example, I would still *experience* myself as *immediately* understanding what you are saying.¹³ Since, as I have discussed earlier, the experiential nature of a mental state is not to be equated with its epistemic complexity, we can assume that DP operates on several subpersonal mechanisms. These are, however, not explicitly represented. Hence it makes sense to describe DP as resulting from *transparent* social cognitive states. By doing so, it is possible to keep its phenomenal status as immediate and direct, while not equating this quality with its epistemic status. In contrast, theorizing and personal-level simulation have a quite different phenomenal characteristic. In these cases, the process of *constructing* a specific insight about the other is part of the experience, may this be by explicitly simulating the person (e.g., “If I was her, what would make me excited about having a cat?”), or through theoretical inference (e.g., “People usually own cats to feel less alone, maybe she is excited to have a furry companion now”). They can thus be said to result from *opaque* social cognitive states. What distinguishes transparent from opaque states is the degree to which one’s own social cognitive processing, which is directed at the other person, is explicitly represented *as* a process.

¹² Note that Schilling and Cruse have already used the abbreviation “1-3sE” to describe levels of situated embodiment. I thus chose a lower case “s” to emphasize the difference (cf. Schilling & Cruse 2008, p. 72).

¹³ “There is, so to speak, nothing that gets in the way, and it is not as if I am first directed at an intermediary, something different from the state, and then only in a secondary step target it.” (Zahavi 2011, p. 548).

However, as already mentioned, I see the need to modify Metzinger's conception of 3E in order to reflect a proper distinction between transparent and opaque social states. I claim that opaque states exhibit an additional level of representation, since the representation process itself is part of the phenomenal experience. In order to emphasize that this is a special and probably rare phenomenon, I introduce the level of "3sE+". Both transparent and opaque social states are certainly to be located at the third level of embodiment, since they possess phenomenal properties. Metzinger suggests that the distinctive feature of 3E in contrast to lower levels is that it enables the system to identify itself with its body (cf. Metzinger 2014, p. 274). The resulting phenomenal properties of self-identification and selfhood stem from the experienced immediacy that comes with transparency (cf. *ibid.*, p. 273). If this is the case, it can be assumed that phenomenal states are not *either* transparent *or* opaque, but that transparency is part of *any* phenomenal state. The degree to which the representation process is explicitly represented varies, transparency and opacity are thus gradually arising properties (cf. Metzinger 2003, p. 358). Additionally, it could well be that there is a constant oscillation between transparency and opacity, depending – for example – on specific contexts and situations. However, opacity and the resulting experiences seem to be more high-level features that can only be found in a small subgroup of species. This is obvious in social understanding, since full-fledged theoretical inference and high-level simulation are not very likely to be found in most non-human animals and human infants. It seems that in the case of opaque states there is an additional level of representation that requires a higher level of sophistication, which should be made more explicit in the hierarchical framework. Transparent and opaque mental states – at least in this case for social understanding – reflect two different kinds of phenomenal experiences that might also have different underlying mechanisms. I thus introduce, in order to do justice to this difference, an additional level of 3sE, namely 3sE+. 3sE+ describes those phenomenal states during which one is aware of the con-

structing process and which occurs in situations that require this kind of reasoning in order to disambiguate the input. This additional distinction at the level of 3sE enables a more detailed view and underlines the difference between transparency and opacity.

One question that arises at this point is the following. We have assumed that opacity means to phenomenally represent (parts of) the actual process of representation. Does that mean that in the case of theorizing and simulation one would find their underlying representational processes to be subpersonal kinds of theoretical inference and simulation? There are two points that speak against this assumption. First, there are justified worries that the conception of implicit theorizing as an unconscious process stretches the concept of a theory too far (e.g., Blackburn 1992). These arguments against TT have been presented extensively in the literature and I will thus not repeat them here. In the case of simulation, secondly, it seems that subpersonal or low-level simulation does not necessarily generate the phenomenal experience of simulating. Consider the many studies that have been conducted to explore whether the activity of the mirror neuron system can be seen as a kind of implicit simulation that enables social understanding (for a review, see for example Cataneo & Rizzolatti 2009). In most of these experiments that found mirror neuron activity to be correlated with social understanding, it seems that the phenomenal experience has the character of DP rather than explicit simulation.¹⁴ Such a view, as I hope to have shown, has two advantages. It describes different kinds of phenomenal experiences in social encounters and distinguishes them by referring to the concepts of transparency and opacity.

5.2 Second-order social embodiment (2sE)

Assuming that there is something like a representational body model, we can now ask which

¹⁴ Note that this is a speculative claim, since almost none of the studies contain phenomenological reports. It could be fruitful, however, for future research to pay more attention to the experience that participants have in a specific experimental setting. This would help to understand which kind of epistemic mechanism generates which kind of experience.

parts of it can be exploited for social cognition. In order to do so, let me briefly recapitulate how to conceive of this body model. It has been described as a “grounded, predictive body model that continuously filters data in accordance with geometrical, kinematic and dynamic boundary conditions” (Metzinger 2014, p. 273). Furthermore, Metzinger predicts that parts of this model can be shared by individuals: “[...] on a certain level of functional granularity, this type of core representation [i.e., the body model] might also describe the generic, universal geometry which is shared by all members of a biological species” (*ibid.*, p. 273; see also Schilling & Cruse 2012). Together with Gallese he argues elsewhere that the mirror neuron system plays a crucial role in generating a basis for both an “internal model of reality” as well as a “shared action ontology” (Metzinger & Gallese 2003, p. 550). This means, as I take it, that the body model contains information that represents one’s own body, but is not completely self-specific. To see this, consider that in order to be shared, representations must not be too specific as to not generalize to the bodies of others. I will come back to this point soon. This consequence worried Newen, leading him to reject the view that mirror neurons form a basis for social cognition:

Why are mirror neurons not an essential part of understanding others? They represent a type of action or emotion that is independent from a first- or third-person perspective; but the distinction between self and other is an essential part of understanding others ([this collection](#), p. 4).

This raises the question of what exactly it is that can be shared by individuals. Since these considerations are central to the possibility of exploiting the body model for social understanding, I now aim to refute the worry and give a possible answer to the question.

Mirror neurons were discovered in the premotor cortex of macaque monkeys more than 20 years ago. They fire, as is famously known, both when an individual executes and observes an action (Gallese et al. 1996; Rizzolatti et al. 1996;

Rizzolatti & Craighero 2004). Although there is considerable controversy about their existence in humans (Hickok 2009), their actual function (Jacob 2008), and their explanatory power (Borg 2007), they are considered by many researchers to form one of the crucial systems for understanding others (e.g., Stanley & Adolphs 2013, p. 512). Mirror neurons are indeed neutral to the agent of an action – they fire whether an action is executed by oneself or another person. Insofar, critics are right to say that it is not obvious how they could provide the important distinction between self and other. However, it seems that there are two important facts left out in this line of thinking. Firstly, it has been suggested that there are inhibition mechanisms that “control” shared representations and provide the basis for a self-other distinction (for a more detailed discussion, see Brass et al. 2009). Secondly, mirror neurons have always been presented as being embedded in a *system* (hence mirror neuron system, e.g., Cattaneo & Rizzolatti 2009; Iacoboni & Dapretto 2006; Rizzolatti & Craighero 2004). This system consists of areas which contain mirror neurons, but also regions which contain neurons that do not have bimodal properties and encode only self-generated actions, as described by Jeannerod & Pacherie (cf. 2004, pp. 131–132).¹⁵ Thus, it is correct that mirror neurons *alone* do not distinguish between self and other. However, this is a rather impoverished view, since they should never be considered in isolation. A similar thought which helps to refute the worry is given by De Vignemont who adopts the view that mirroring can be seen as sharing body repres-

¹⁵ “The problem of agent-identification, however, is solved by the fact that other premotor neurons (the canonical neurons) and, presumably many other neuron populations as well, fire only when the monkey performs the action and not when it observes it performed by another agent. This is indeed another critical feature of the shared representations concept: they overlap only partially, and the non-overlapping part of a given representation can be the cue for attributing the action to the self or to the other. The same mechanism operates in humans. Neuroimaging experiments where brain activity was compared during different types of simulated actions (e.g., intending actions and preparing for execution, imagining actions, observing actions performed by other people) revealed, first, that there exists a cortical network common to all conditions, to which the inferior parietal lobule (areas 39 and 40), the ventral premotor area (ventral area 6), and part of SMA contribute; and second, that motor representations for each individual condition are clearly specified by the activation of cortical zones which do not overlap between conditions [...]” (Jeannerod & Pacherie 2004, pp. 131–132)

entations (2014a). She argues that shared body representations do not threaten a self-other distinction because they always contain information that is too self-specific to be shared. They are, in her words, “[...] Janus-faced. They face inward as representations of one’s body and they face outward as representations of other people’s bodies” (De Vignemont 2014b, p. 135).

A closer look at her conception also yields a possible answer to the question of what it is that can be shared with others. De Vignemont argues that it must be a rather coarse-grained representation of one’s body, since bodies differ considerably in many aspects like size, gender, posture etc. This representation, which De Vignemont dubs the “body map” (De Vignemont 2014a, p. 289, 2014b, p. 134), contains information about the basic configuration of body parts and thus serves as a functional tool to localize bodily experiences. Irrespective of individual differences of this map, some of its content is so coarse-grained that humans are still able to imitate others or experience vicarious bodily sensations, both of which have been claimed to draw on shared body representations. In other words, what can be shared is that part of the body map whose content is general enough to apply to all kinds of bodies, no matter their differences.

Although this is surely no exhaustive inquiry of the matter, these thoughts provide an idea of how to view 2sE as enabling social cognition: at the representational level, there are parts of the body model which can be shared with others.¹⁶ These parts, however, have to be embedded in a system that also contains self-specific information. Otherwise it would be impossible to attribute an action, an experience or observation to the agent concerned. It now becomes obvious why I claimed earlier that a self-other distinction does not need a *phenomenal* representation of one’s body. The unconscious body model and its shared parts seem well furnished to provide such a distinction and thus make unconscious social processes such as mimicry and involuntary imitation possible.

5.3 First-order social embodiment (1sE)

Although interaction is certainly a topic that has been the least explored by researchers of social cognition, it nevertheless should be considered carefully by any theory that aims to provide a comprehensive view on social understanding. Including interaction is particularly challenging, since most attempts to do so came from proponents of an enactive perspective on the mind. However, I have argued that a pluralistic model of social cognition cannot simply combine enactive claims with cognitive ones (see section 3 “Multiplicity needs coherence”). What is needed is an approach of social understanding that integrates interaction as a phenomenon that most probably does not need explicit, high-level representation. 1sE offers a way to describe such low-level social processes. Knoblich and Sebanz, for example, review several cases of “social coupling”. Individuals tend to synchronize their movements if they are sitting next to each other in a rocking chair (cf. Knoblich & Sebanz 2008, p. 2022), a process which can plausibly be described without representation. This sort of “entrainment” (*ibid.*, p. 2023) is a case of coupling during which individuals influence each other’s behavior without consciously intending to do so. There are also cases in the animal kingdom that can be described at the level of 1sE, such as the formation and synchronization of fireflies (Suda et al. 2006).

The next step is to depict the implementation of specifically “social parts” of the body model. What physically grounds them is described at the level of 1sE. One buzzword in the research field of social cognitive neuroscience is “the social brain” (e.g., Dunbar 1998; Gazzaniga 1985). This term refers to all the different areas in the brain that have been found to be correlated to cognitive processing in social situations, including, of course, the mirror neuron system. While the investigation of brain regions and their functions for social cognition is a well-established endeavor, it will be more interesting to look at other possibilities of implementing social cognition. The role of interaction for social cognition, for example, has been hotly disputed in the research field. As I have il-

¹⁶ Sharing means that representational content overlaps, at least partially. For a more detailed discussion on sharing, see De Vignemont 2014b; Jeannerod & Pacherie 2004.

lustrated earlier, some claim that interaction dynamics could *constitute* social cognitive mechanisms (De Jaegher et al. 2010). However, such a view is only sustainable in a radically enactive set of assumptions and as such is not an option for the framework I am suggesting here. What should be considered, though, is whether being in an interaction is *necessary* for some social cognitive states. It has been suggested by recent studies that activation patterns differ depending on the situational context and the degree of emotional engagement in a social situation (Schilbach et al. 2013). These results point to this possibility, but it still needs more careful investigation whether or not they justify the claim that interaction in any way *physically grounds or enables* social cognition.

Such basic and non-representational forms of social understanding have been neglected by the research field for a long time and are in need of more empirical and philosophical investigation. Especially research on joint action and coupled systems is therefore important to sort out 1sE.

6 Conclusion

My first goal in this commentary was to show that MV as a pluralistic view on social understanding is a valuable contribution to the interactive turn. It has the potential to integrate insights from different directions of empirical and theoretical research and thus to yield a comprehensive account on social cognition. However, I argued that such an approach needs careful consideration concerning its metaphysical background assumptions. I demonstrated that parts of MV as laid out by Newen are not fully compatible and that it thus needs a different kind of framework which allows a coherent picture.

I presented an alternative model by applying Metzinger's framework of 1-3E to social cognition, hence 1-3sE. Although the details are still to be spelled out in future research, 1-3sE has several advantages that enable a coherent and fruitful framing of MV. It integrates all four social mechanisms mentioned by Newen and thus can be seen as a pluralistic account of social cognition. What is different,

however, is that those four elements are described at different levels of description. As such they all play a specific role in the overall image of social understanding and merge into a manifold, but unified picture. Basic interaction, in this theory, can be accounted for without making radical claims in either direction; we do not need to assume that the mind is relational, as claimed by proponents of the enactive theory. However, we also do not have to ascribe a high level of sophistication to a system in order to be able to interact. In my proposal, interaction (or at least simple interactive mechanisms) can function without any complex representation. Interaction is thus located at the lowest level of the hierarchy, namely 1sE. The next level of social embodiment describes representational and computational processes that subserve social cognition. I showed in which ways a model of one's own body could enable social cognition and which parts of such a model could possibly be shared with others. 2sE encompasses these processes. I further argued that DP should be treated as a phenomenological rather than epistemological concept and should thus be described at the level of 3sE. By doing so, I aimed to avoid mixing up different levels of description and to yield a coherent usage of the term. High-level simulation and theoretical inference have been described at the level of 3sE+, the highest level of the hierarchy, thus doing justice to the fact that they are very special and probably rare cases of social cognition. The application of the notions of transparency and opacity offered a way to emphasize the phenomenological variety that comes with different social situations.

There are still many open questions and this is by no means an exhaustive description of how 1-3sE can be used to frame social understanding. My goal here was to highlight its potential to provide a framework which offers novel ways to (1) incorporate the phenomenal level of description with its representational counterparts, (2) to integrate the role of the body as shaping and grounding social cognitive processes and thus (3) to depict social cognition as a representational, but still embodied ability.

Acknowledgements

I would like to thank the Barbara-Wengeler-Stiftung for their financial support, Thomas Metzinger and Jennifer Windt for giving me the opportunity to be part of this project, two anonymous reviewers for their valuable feedback, and Luke Miller for helpful suggestions on form and content.

References

- Adolphs, R. (2006). How do we know the minds of others? Domain-specificity, simulation, and enactive social cognition. *Brain Research*, 1079 (1), 25-35.
[10.1016/j.brainres.2005.12.127](https://doi.org/10.1016/j.brainres.2005.12.127)
- Blackburn, S. (1992). Theory, observation and drama. *Mind & Language*, 7 (1-2), 187-230.
[10.1111/j.1468-0017.1992.tb00204.x](https://doi.org/10.1111/j.1468-0017.1992.tb00204.x)
- Borg, E. (2007). If mirror neurons are the answer, what was the question? *Journal of Consciousness Studies*, 14 (8), 5-19.
- Brass, M., Ruby, P. & Spengler, S. (2009). Inhibition of imitative behaviour and social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364 (1528), 2359-2367.
- Bruin, L. C. de & Kästner, L. (2012). Dynamic embodied cognition. *Phenomenology and the Cognitive Sciences*, 11 (4), 541-563. [10.1007/s11097-011-9223-1](https://doi.org/10.1007/s11097-011-9223-1)
- Cattaneo, L. & Rizzolatti, G. (2009). The mirror neuron system. *Archives of neurology*, 66 (5), 557-560.
- Clark, A. (2013a). Dreaming the whole cat: Generative models, predictive processing, and the enactivist conception of perceptual experience. *Mind*, 121 (483), 753-771.
- (2013b). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36, 181-253.
[10.1017/S0140525X12000477](https://doi.org/10.1017/S0140525X12000477)
- (2015). Embodied prediction. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Cruse, H. & Schilling, M. (2015). Mental states as emergent properties - From walking to consciousness. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- De Jaegher, H., Di Paolo, E. & Gallagher, S. (2010). Can social interaction constitute social cognition? *Trends in Cognitive Sciences*, 14 (10), 441-447.
[10.1016/j.tics.2010.06.009](https://doi.org/10.1016/j.tics.2010.06.009)
- De Jaegher, H. & Paolo, E. (2007). Participatory sense-making. *Phenomenology and the Cognitive Sciences*, 6, 485-507. [10.1007/s11097-007-9076-9](https://doi.org/10.1007/s11097-007-9076-9)
- De Vignemont, F. (2010). Knowing other people's mental states as if they were one's own. In D. Schmicking & S. Gallagher (Eds.) *Handbook of Phenomenology and Cognitive Science* (pp. 283-299). Dordrecht, NL: Springer.
- (2014a). Shared body representations and the 'whose' system. *Neuropsychologia*, 55, 128-136.
[10.1016/j.neuropsychologia.2013.08.013](https://doi.org/10.1016/j.neuropsychologia.2013.08.013)

- (2014b). Acting for bodily awareness. In R. Shapiro (Ed.) *The Routledge Handbook of Embodied Cognition* (pp. 287-295). New York, NY: Routledge.
- Di Paolo, E. & Thompson, E. (2014). The enactive approach. In R. Shapiro (Ed.) *The Routledge Handbook of Embodied Cognition* (pp. 68-78). New York, NY: Routledge.
- Dunbar, R. I. (1998). The social brain hypothesis. *Evolutionary Anthropology*, 178-190.
- Engel, A. K., Maye, A., Kurthen, M. & König, P. (2013). Where's the action? The pragmatic turn in cognitive science. *Trends in Cognitive Sciences*, 17 (5), 202-209.
- Fiebich, A. & Coltheart, M. (in press). Varieties of social understanding. *Mind & Language*
- Fuchs, T. & Jaegher, H. (2009). Enactive intersubjectivity: Participatory sense-making and mutual incorporation. *Phenomenology and the Cognitive Sciences*, 8 (4), 465-486.
- Furlanetto, T., Cavallo, A., Manera, V., Tversky, B. & Becchio, C. (2013). Through your eyes: Incongruence of gaze and action increases spontaneous perspective taking. *Frontiers in Human Neuroscience*, 7.
- Gallagher, S. (2001). The practice of mind: Theory, simulation or primary interaction? *Journal of Consciousness Studies*, 8 (5-7), 83-108.
- (2008). Direct perception in the intersubjective context. *Consciousness and Cognition*, 17 (2), 535-543. [10.1016/j.concog.2008.03.003](https://doi.org/10.1016/j.concog.2008.03.003)
- Gallagher, S., Hutto, D. D., Slaby, J. & Cole, J. (2013). The brain as part of an enactive system. *Behavioral and Brain Sciences*, 36 (4), 421-422. [10.1017/S0140525X12002105](https://doi.org/10.1017/S0140525X12002105)
- Gallagher, S. & Hutto, D. D. (2008). Understanding others through primary intersubjectivity and narrative practice. In J. Zlatev, C. Shina & E. Itkonen (Eds.) *The Shared Mind: Perspectives on Intersubjectivity* (pp. 1-18). Amsterdam, NL: John Benjamins.
- Gallese, V., Fadiga, L., Fogassi, L. & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, 119, 592-609.
- Gallotti, M. & Frith, C. D. (2013). Social cognition in the we-mode. *Trends in Cognitive Sciences*, 17 (4), 160-165. [10.1016/j.tics.2013.02.002](https://doi.org/10.1016/j.tics.2013.02.002)
- Gazzaniga, M. S. (1985). *The social brain: Discovering the networks of the mind*. New York, NY: Basic Books.
- Gibson, J. (1977). The theory of affordances. In R. Shaw (Ed.) *Perceiving, acting, and knowing. Toward an ecological psychology* (pp. 67-82). Hillsdale, NJ: Erlbaum.
- (1979). *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin Company.
- Goldman, A. I. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. New York, NY: Oxford University Press.
- Gopnik, A. & Meltzoff, A. N. (1997). *Words, thoughts, and theories. Learning, development, and conceptual change*. Cambridge, MA: MIT Press.
- Hickok, G. (2009). Eight problems for the mirror neuron theory of action understanding in monkeys and humans. *Journal of Cognitive Neuroscience*, 21 (7), 1229-1243. [10.1162/jocn.2009.21189](https://doi.org/10.1162/jocn.2009.21189)
- Hohwy, J. (2013). *The predictive mind*. Oxford, UK: Oxford University Press.
- (2014). The self-evidencing brain. *Nous*. [10.1111/nous.12062](https://doi.org/10.1111/nous.12062)
- (2015). The neural organ explains the mind. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Hutto, D. D. (2008). The narrative practice hypothesis: Clarifications and implications. *Philosophical Explorations*, 11 (3), 175-192. [10.1080/13869790802245679](https://doi.org/10.1080/13869790802245679)
- Iacoboni, M. & Dapretto, M. (2006). The mirror neuron system and the consequences of its dysfunction. *Nature Reviews Neuroscience*, 7 (12), 942-951. [10.1038/nrn2024](https://doi.org/10.1038/nrn2024)
- Jacob, P. (2008). What do mirror neurons contribute to human social cognition? *Mind Language*, 23 (2), 190-223. [10.1111/j.1468-0017.2007.00337.x](https://doi.org/10.1111/j.1468-0017.2007.00337.x)
- Jeannerod, M. & Pacherie, E. (2004). Agency, simulation and self-identification. *Mind and Language*, 19 (2), 113-146. [10.1111/j.1468-0017.2004.00251.x](https://doi.org/10.1111/j.1468-0017.2004.00251.x)
- Kilner, J. M., Friston, K. J. & Frith, C. D. (2007). Predictive coding: an account of the mirror neuron system. *Cognitive Processing*, 8 (3), 159-166.
- Knoblich, G. & Sebanz, N. (2008). Evolving intentions for social interaction: From entrainment to joint action. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363 (1499), 2021-2031. [10.1098/rstb.2008.0006](https://doi.org/10.1098/rstb.2008.0006)
- Longo, M. R. & Tsakiris, M. (2013). Merging second-person and first-person neuroscience. *Behavioral and Brain Sciences*, 36 (04), 429-430.
- Metzinger, T. (2003). Phenomenal transparency and cognitive self-reference. *Phenomenology and the Cognitive Sciences*, 2, 353-393.
- (2004). *Being no one: The self-model theory of subjectivity*. Cambridge, MA: MIT Press.
- (2006). Different conceptions of embodiment. *Psyche*, 12 (4)
- (2007). Self models. *Scholarpedia*, 2 (10), 4174.

- (2014). First-order embodiment, second-order embodiment, third-order embodiment: From spatiotemporal self-location to minimal selfhood. In R. Shapiro (Ed.) *The Routledge Handbook of Embodied Cognition* (pp. 272-286). New York, NY: Routledge.
- Metzinger, T. & Gallese, V. (2003). The emergence of a shared action ontology: Building blocks for a theory. *Consciousness and Cognition*, 12 (4), 549-571. [10.1016/S1053-8100\(03\)00072-2](https://doi.org/10.1016/S1053-8100(03)00072-2)
- Newen, A. (2015). Understanding others - The person model theory. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Overgaard, S. & Michael, J. (2013). The interactive turn in social cognition research: A critique. *Philosophical Psychology*, 28 (2), 1-25. [10.1080/09515089.2013.827109](https://doi.org/10.1080/09515089.2013.827109)
- Pezzulo, G., Barsalou, L. W., Cangelosi, A., Fischer, M. H., McRae, K. & Spivey, M. J. (2013). Computational grounded cognition: A new alliance between grounded cognition and computational modeling. *Frontiers in Psychology*, 3.
- Rizzolatti, G. & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, 27 (1), 169-192. [10.1146/annurev.neuro.27.070203.144230](https://doi.org/10.1146/annurev.neuro.27.070203.144230)
- Rizzolatti, G., Fadiga, L., Gallese, V. & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3 (2), 131-141.
- Rowlands, M. (2009). Enactivism and the extended mind. *Topoi*, 28 (1), 53-62. [10.1007/s11245-008-9046-z](https://doi.org/10.1007/s11245-008-9046-z)
- Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T. & Vogeley, K. (2013). Toward a second-person neuroscience. *Behavioral and Brain Sciences*, 36 (4), 393-414. [10.1017/S0140525X12000660](https://doi.org/10.1017/S0140525X12000660)
- Schilling, M. & Cruse, H. (2008). The evolution of cognition: From first order to second order embodiment. In I. Wachsmuth & G. Knoblich (Eds.) *Lecture notes in computer science Lecture notes in artificial intelligence: Vol. 4930. Modeling communication with robots and virtual humans. Second ZiF Research Group International Workshop on Embodied Communication in Humans and Machines, Bielefeld, Germany, April 5 - 8, 2006; revised selected papers* (pp. 77-108). Berlin: Springer.
- (2012). What's next: Recruitment of a grounded predictive body model for planning a robot's actions. *Frontiers in Psychology*, 3.
- Seth, A. K. (2015). Inference to the best prediction. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Stanley, D. A. & Adolphs, R. (2013). Toward a neural basis for social behavior. *Neuron*, 80 (3), 816-826. [10.1016/j.neuron.2013.10.038](https://doi.org/10.1016/j.neuron.2013.10.038)
- Suda, T., Tschudin, C., Tyrrell, A., Auer, G. & Bettstetter, C. (2006). *Fireflies as role models for synchronization in ad hoc networks. In Proceedings of the 1st international conference on Bio inspired models of network, information and computing systems*. New York, NY: ACM.
- Thompson, E. (2010). *Mind in life: Biology, phenomenology, and the sciences of mind*. Cambridge, MA, London: Belknap.
- Varela, F. J., Thompson, E. & Rosch, E. (1993). *The embodied mind: Cognitive science and human experience*. Cambridge, MA: MIT Press.
- Zahavi, D. (2011). Empathy and direct social perception: A phenomenological proposal. *Review of Philosophy and Psychology*, 2 (3), 541-558. [10.1007/s13164-011-0070-3](https://doi.org/10.1007/s13164-011-0070-3)
- Zimmer, D. E. (1989). Wilde Kinder. In D. E. Zimmer (Ed.) *Experimente des Lebens* (pp. 21-47). Zürich: Haffmanns Verlag.

A Multiplicity View for Social Cognition: Defending a Coherent Framework

A Reply to Lisa Quadt

Albert Newen

Lisa Quadt's commentary focuses on my theory about the multiple epistemic strategies humans use to receive information about one other's mental phenomena. She develops a principle worry about the theory's underlying metaphysical foundations, arguing that I am committed to an incoherent metaphysical framework. In this reply, I show that I am not committed to the position she attributes to me and I outline an alternative framework that is my actual background view. I illustrate this framework by discussing emotions and argue that emotions are individuated as integrated patterns of characteristic features. This enables me to combine a representational account of emotions with a theory of direct perception of basic emotions as well as with an understanding of some emotions relying on theory-based inferences. Thus, I have a coherent metaphysics. Finally, I show that the alternative suggested by Quadt has its own problems.

Keywords

Direct perception | Metaphysical foundation | Person model theory | Social cognition | Transparency

Author

Albert Newen

albert.newen@rub.de

Ruhr-Universität Bochum
Bochum, Germany

Commentator

Lisa Quadt

lisquadt@students.uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

With my PMT (person model theory), I aim to answer two questions. While the first question asks which epistemic strategy humans use to access the mental states of others and to gather information about them, the second question asks how the information we obtain to understand others is stored and organized. The answer to the second question is the core of the PMT. It states that information about other in-

dividuals or types of persons is stored and organized in person models and that these are realized on two levels, i.e. the implicit level of person schemata and the explicit level of person images. It further argues that philosophical theories so far have predominately ignored the fact that we usually understand others relying on rich background information concerning them and their situation.

Lisa Quadt's commentary focuses on my theory concerning the epistemic strategies humans use to receive information about others' mental phenomena, and she develops a principle worry about the underlying metaphysical foundations. I am grateful for this challenge, which gives me the opportunity to clarify my background view. The MV (multiplicity view) outlined in the target paper claims that we do not rely on one epistemic strategy alone, as is suggested by most proposals in the literature, but that we rely on a multiplicity of strategies which, for the most part, are implicitly activated on the basis of contextual conditions. These strategies include simulation strategies, theory-based inferences, and direct perception as well as understanding by social interaction and by relying on narratives. Quadt's main worry is that MV may be based on an incoherent metaphysics and is thus unacceptable as it stands. In the first part of her reply she aims to defend the incoherence claim, while in the second part she offers an alternative metaphysical framework. My reply is structured as follows: In the next paragraph I briefly describe how Quadt defends her claim about the supposed incoherence of my metaphysical background and show that I am not committed to the incoherent framework she attributes to me. In the second section, I make explicit my actual background metaphysics (which was not the focus of my article) and argue that it is coherent, reinforcing that I am not committed to the metaphysics that Quadt attributes to my position. Finally, I argue that the alternative metaphysics suggested by Quadt relies on a distinction between transparency and opacity that cannot carry the weight it is supposed to carry.

2 Am I committed to an incoherent metaphysics?

Quadt describes correctly that the MV I advocate combines epistemic strategies that are described in several different positions, including ST (Simulation Theory) (Goldman 2006), TT (Theory-Theory) (e.g., Gopnik & Meltzoff 1997), and IT (Interaction Theory) (Gallagher 2001), as well as theory of direct perception

(Gallagher 2008). As a consequence, she presupposes that I am committed to the metaphysical foundations of each of these positions, while each position argues for a distinct epistemic strategy. If I were committed to accepting such metaphysical foundations, I would thereby offer an incoherent metaphysics. Quadt shows this by arguing that Simulation Theory and Theory-Theory, on the one hand, presuppose metaphysical claims that are not consistent with the presuppositions from Theories of DP (direct perception) and ITs, on the other hand (3). Quadt claims that ST and TT are *cognitivist theories* that presuppose internalism, mental representations, and the idea that mental phenomena are private hidden entities to which we have no direct access. To register mental phenomena we have to rely on perceiving the behaviour and expressions of other people and have to *infer* the existence of mental phenomena. Quite the opposite view is taken by the *non-cognitivist theories* of DP and IT. They allow for externalism of mental phenomena (as being realized by two people and their interaction), they deny the existence of mental representations, and they presuppose that mental phenomena are not hidden but directly perceivable. Thus they rely on non-inferential access to mental phenomena by direct perception. The following quote illustrates the main features of the contrast Quadt develops:

The difference between cognitivist and non-cognitivist pictures of social cognition, in the cases that I just described, seems to boil down to the metaphysical assumption of whether or not there are hidden cause in the outside world that require an inference or representational mechanism in order to access and process them. While ST and TT clearly assume such a view, DP denies it. Therefore, I claim that MV cannot simply combine theoretical elements that draw on such considerable metaphysical differences. (Quadt 2015, p. 5)

My first general reply to this worry is that I only take on the description of an *epistemic strategy* of acquiring and using information

about other people in order to understand them. An epistemic strategy like a simulation (to put oneself in the other person's shoes) or a theory-based inference is not automatically connected to a metaphysical commitment. De facto, the philosophers who are famous for holding ST or TT combine their view with a metaphysical background, but it does not follow that the epistemic strategy they describe *must be combined* with the metaphysical background they offer. We can easily see this for example in the case of two epistemic strategies like theory-based inferences and direct perception of mental phenomena. These can be easily combined in a way that allows that some mental phenomena with intense expressive components like basic emotions (Ekman et al. 1972) can be directly perceived (see below), while complex mental phenomena like propositional attitudes may be at least often inferred if the social understanding cannot rely on honest utterances but only on some ambiguous behavioural cues. Thus, the de facto incompatibility of the metaphysical presuppositions of the two main lines of theories of social understanding does not imply that I am committed to inheriting both presuppositions and that I thus run into an incoherent metaphysics. In fact, I do not presuppose two metaphysical principles for the same mental phenomenon; instead I only need to allow for the application of two epistemic strategies of understanding mental phenomena, which may be applied to different mental phenomena (or to the same type of mental phenomenon in different situations). In the next section I outline my alternative metaphysics and illustrate both that it is coherent and that it can allow for direct perception as one epistemic strategy for registering some mental phenomena.

3 Defending direct perception in an alternative metaphysical framework

In general, I prefer to think of mental phenomena as representational, but I do not see that this prevents me from integrating the epistemic strategy of direct perception. Furthermore, I characterize basic emotions as realized in one individual (individualism but not internalism).

At the same time, I remain neutral as to whether joint emotions (e.g. joint enthusiasm about a goal achieved by one's team) have to be analysed as extended emotions. Furthermore, I think that basic emotions are not hidden mental phenomena but can be directly perceived e.g. on the basis of face-based recognition of emotions. Thus, I think that some mental phenomena can be registered non-inferentially. But of course, direct perception of some mental phenomena is *only one* of at least four epistemic strategies that we can use, depending on the context.

To sketch my theory of direct perception I will focus on basic emotions like anger, fear, happiness, sadness, etc. (for a classification of emotions see Zinck & Newen 2008). My metaphysical view of emotional episodes is that they are integrated patterns of characteristic features (Welpinghus & Newen 2012; Newen et al. 2015). Let me use the example of fear as illustrated in Newen et al. (2015): an emotional episode of fear towards an aggressive dog is constituted by the integration of the following characteristic features: (1) a typical physiological arousal that is a consequence of bodily changes due to changes in the autonomic nervous system, including increased heart rate and flat breathing; (2) a typical behavior or behavioral disposition, including flight or freezing behavior; (3) a typical facial expression, gesture, or body posture, etc.; (4) a typical phenomenal experience of fear; (5) a typical (explicit) cognitive evaluation of the dog in front of me (e.g., "This is an aggressive pit bull"). Furthermore, every emotional episode has (6) an intentional object, i.e. the dog in front of me. Features 1–5 are integrated into an (often implicit) appraisal of the intentional object as dangerous. The emotional episode is constituted by the integration of all the characteristic features mentioned so far, including the appraisal. This view allows that in another implementation *some* features would be missing. For example, the explicit cognitive evaluation of the dog as an aggressive pit bull is not necessary to be in fear towards the dog in front of me. Or the facial expression may be inhibited, due to intense training to attain a poker face, yet I may still be in fear. As long as a minimum of features is realized, we still have

an episode of fear. The two main features that are necessary in all emotional episodes are a registration of minimal physiological arousal and an intentional object. The integration of both is needed to have an emotional episode (Barlassina & Newen 2013). But other features may be lacking while still remaining characteristic of most episodes of the relevant type of emotion. One might wonder why I do not include neural correlates. Since I argue from a position of antecedent naturalism, neural correlates are not an extra component in addition to the characteristic features already mentioned above. We might mention neural correlates as an informative aspect for the individuation of certain features of emotion, but we do not have to, since they concern the same features that have already been mentioned, with information accessed in a different manner.

If one accepts the ontology of emotions as individuated by an integrated pattern of characteristic features, it follows that the expression of an emotion by face, body posture, and gestures is a *constitutive* part of the emotional episode (and not a causal consequence). Thus, I do not hold internalism about mental phenomena. Given this theory of the individuation of emotions, I also argue for the thesis that one way of recognizing emotions is by perceiving the relevant pattern (Newen et al. 2015). A recognition of the other person's fear can be attained by directly perceiving the pattern of fear. How can we account for this, while at the same time accepting that the feeling of fear is a private subjective experience in so far as a person still may have the feeling even if she is able to keep a poker face? Perceiving fear is comparable to perceiving a house. Both are processes of pattern recognition on the basis of a minimal package of characteristic features: I can recognize a drawing as one of a house, even if one or two of the characteristic features of a house are missing. How is this possible? Perceiving an object is not a purely passive process, like taking a photograph; it is a constructive process.¹ One

important aspect of the constructive process is the enrichment of selected core sensory information. And one way of realizing this enrichment is by the activation of a rich memorized mental image that best suits the core sensory information. If we have learned the relevant pattern of what a house looks like from the outside, and memorized a respective mental image, then seeing a child's drawing initiates an interaction of bottom-up and top-down processes. These include the activation of this stored mental image, such that it enriches the core sensory information to form a perceptual experience of seeing a drawing of a house even if the front door is missing in the drawing.

The same process of pattern recognition takes place in the case of recognizing an emotion like fear. The relevant pattern of fear is formed either on the basis of having personally experienced a situation of fear or on the basis of having observed others in such situations. One thereby acquires a memorized pattern of fear with typical features. If one now observes a person with a typical facial expression in a situation where she is being attacked by a dog, one can see the fear of the person. The perception of fear is realized by seeing the freezing behaviour, the facial expression, and the intentional object (i.e. the aggressive dog), because these features activate as part of the process of perceptual processing the whole pattern of fear. Thus, I can *perceive* fear in the face of the person being attacked by the dog. The theory of perception is one according to which perceptual processing allows for a systematic enrichment of information and for influencing of perceptual processes by memorized images or background knowledge. These top-down influences are discussed under the label cognitive penetration. So I am committed to the view of perception as cognitively penetrated as it is defended in detail in Vetter & Newen (2014). But this does not involve any claims concerning the metaphysical commitments ascribed to me by Quadts in her commentary. Recognition of emotions is analysed in a framework that explicitly allows for mental representations but specifies them in a way that nevertheless allows for direct perception as one form of access to the recognition of

1 All modern theories of perception account for this constructive component, e.g. O'Regan's and Noë's theories of enacted perception (O'Regan & Noë 2001; Noë 2005), as do theories of cognitive penetration (Macpherson 2012; Siegel 2012) and theories of predictive coding (Hohwy 2013; see also Hohwy this collection; Clark this collection).

emotions. As has been spelled out in detail elsewhere (see [Newen et al. 2015](#)), in principle I allow for three types of recognizing of emotions: two types of direct perception are distinguished in terms of top-down processes of shaping perception involving background images or beliefs; and one is characterized by theory-based inferences. Thus, I distinguish “(1) (a basic form of) perceiving an emotion in the (near) absence of any top-down processes, and (2) perceiving an emotion in a way that significantly involves some top-down processes (a strongly concept-modified form of perception). Both types of perceiving emotions can be distinguished from (3) inference-based evaluation of an emotion pattern. The latter presupposes a stable evaluation of an emotion as being *F*, which then may be modified or reevaluated by reflecting on the information” ([Newen et al. 2015](#), p. 197). To sum up: Direct perception can be based on a metaphysical framework that regards emotions as integrated patterns of characteristic features and this allows me to combine it with presupposing mental representations of emotions (as memorized rich patterns), on the one hand, as well as with a non-inferential recognition of some emotional episodes on the other. The pattern theory of emotion is furthermore able to account for internalistic features of emotions like the feeling of fear, but also for individualistic yet expressive features like behavior and expression in face, gesture, and body posture. This metaphysics of emotions is coherent and is compatible with several epistemic strategies for recognizing them, e.g. direct perception as well as theory-based inferential understanding.

Let me make a further clarificatory remark about my reply to the coherence worry: I illustrated my metaphysics taking emotional episodes as a core example. This does not imply that I analyze *all* mental phenomena in this way. Although I think that some mental phenomena can also be individuated as integrated patterns of characteristic features like self-awareness/self-consciousness (see [Gallagher 2013](#)) or object perception, I remain neutral on the question of how far this analysis can be generalized and about the possibility that some mental phenomena need a different metaphysics

as basis. For this reply it is sufficient to have shown what a concrete paradigmatic example of a coherent metaphysics for emotional episodes looks like, in order to prevent the danger of running into an incoherent metaphysics as a unavoidable consequence of the multiplicity view concerning epistemic strategies of understanding others.²

4 Quadts proposal FOR an alternative metaphysical framework

Although I think I do not need an alternative metaphysics, since I have a coherent one already, I would like to briefly comment on Quadts account. She starts with a remark on embodiment. I do not really see any serious disagreement with my views here. For it is fine by me that phenomenal properties and mental representations in general are realized within the body —and sometimes not only in the brain but within our whole body (see the discussion of emotions). Furthermore, I said that in this reply I leave open whether we need an extended realization basis for some mental representations. Quadts alternative proposal, with which she aims to deliver a new framework for a multiplicity view, introduces different levels of embodiment. One way to read her distinction is that it offers a characterization of different types of representation that unfold during ontogeny. This basic idea is entirely consistent with my work. In other papers I discuss in detail the development of different types of representation in ontogeny ([Newen & Vogeley 2003](#); [Newen & Fiebach 2009](#); [de Bruin & Newen 2012](#)). There are of course differences in how one might form types of representation but discussion of these goes beyond the scope of this reply.

Let me now elaborate on an important point of disagreement. Quadts proposal is based, among other things, on the distinction between transparent and opaque ways of being involved in a mental state. She takes this distinction from [Metzinger \(2003, 2004\)](#). We can illustrate this dis-

² Let me highlight that the multiplicity view of understanding others is only one part of my person model theory and this epistemic aspect is in addition defended and further developed by my former PhD-student Anika Fiebach in the following paper which just appeared: [Fiebach & Coltheart 2015](#).

inction using the example of the mental event of perceiving an apple. This event is transparent if I am only consciously aware of the apple, while it is opaque if I am (also) aware of my mental state of seeing the apple: “[w]hat distinguishes transparent from opaque states is the degree to which one’s own social cognitive processing, which is directed at the other person, is explicitly represented as a process” (Quadt 2015, p. 12). The relevant move is Quadt’s claim that the epistemic access of direct perception in social cognition can be explained by transparency, while the epistemic access of simulation and theory-based inference can be explained by opacity.

Here I think she is on the wrong track. This distinction between transparency and opacity in the case of a mental state of attributing a belief leads to the idea that I am not only aware of the other person having a belief with content *p* but that I am also focussing on being consciously aware of the process of my attributing a belief to the other. The latter can of course happen in case of reflective processes of attributing beliefs; but normally we are in a mode of just using our ability to attribute beliefs automatically, focusing on the other’s belief and its content (not on our own process of attributing it). We normally deal with our mental state of attributing beliefs in a transparent way, contrary to the analysis offered by Quadt. Furthermore, direct perception can also be used opaquely in rare cases of being reflectively aware of guiding images: if I am an experienced chess player, I can perceive the chess board in a way that is best described by cognitive penetration, and in some cases I may be aware of the mental image which guides my perception, i.e. I see a position and know how to act because I consciously memorize the fact that I see exactly the same position I saw in a previously played game. Thus, the distinction between transparency and opacity is not helpful for characterizing the different strategies of epistemic access to another’s mental states.

5 Self-models and person models: how are they related?

Finally let me point out an important question raised by Quadt, namely how are person models

and self-models related to each other? A self-model is a special type of person model, the person model that someone develops of herself. This is also done at the two levels of an implicit self-schema and an explicit self-image. I intend to elaborate on the interaction between self-models and person model of others in future articles, but I completely agree with Quadt when she says that there is bi-directional informational exchange regarding both types of models in humans (which is also indicated in my paper in figure 2, p. 21): “I thus conclude that it should not only be considered how the development of a self-model influences social cognition, but also which role social processes play in forming such a self-model” (Quadt 2015, p. 10). The PMT has potential as a framework for a theory of human self-consciousness.

References

- Barlassina, L. & Newen, A. (2013). The role of bodily perception in emotion: In defense of an impure somatic theory. *Philosophy and Phenomenological Research*, 1-42.
- Clark, A. (2015). Embodied prediction. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- de Bruin, L. & Newen, A. (2012). An association account of false belief understanding. *Cognition*, 123 (2), 240-259. [10.1016/j.cognition.2011.12.016](https://doi.org/10.1016/j.cognition.2011.12.016)
- Ekman, P., Friesen, W. V. & Ellsworth, P. (1972). *Emotion in the Human Face*. Oxford: Pergamon Press.
- Fiebich, A. & Coltheart, M. (2015). Various ways to understand other minds. *Mind and Language*.
- Gallagher, S. (2001). The practice of mind: Theory, simulation, or interaction? *Journal of Consciousness Studies*, 8 (5-7), 83-107.
- (2008). Direct perception in the intersubjective context. *Consciousness and Cognition*, 17 (2), 535-543. [10.1016/j.concog.2008.03.003](https://doi.org/10.1016/j.concog.2008.03.003)
- (2013). A pattern theory of self. *Frontiers in Human Neuroscience*, 7 (443), 1-7.
- Goldman, A. I. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. New York, NY: Oxford University Press.
- Gopnik, A. & Meltzoff, A. N. (1997). *Words, Thoughts, and Theories*. Cambridge, MA: MIT Press.
- Hohwy, J. (2013). *The Predictive Mind*. Oxford: Oxford UP.
- (2015). The neural organ explains the mind. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Macpherson, F. (2012). Cognitive penetration of colour experience. Rethinking the issue in light of an indirect mechanism. *Philosophy and Phenomenological Research*, 84 (1), 24-62. [10.1111/j.1933-1592.2010.00481.x](https://doi.org/10.1111/j.1933-1592.2010.00481.x)
- Metzinger, T. (2003). Phenomenal transparency and cognitive self-reference. *Phenomenology and the Cognitive Sciences*, 2, 353-393.
- (2004). *Being no one: The self-model theory of subjectivity*. Cambridge, MA: MIT Press.
- Newen, A., Welpinghus, A. & Juckel, G. (2015). Emotion recognition as pattern recognition: the relevance of perception. *Mind & Language*, 30 (2), 187-208.
- Newen, A. & Fiebich, A. (2009). A developmental theory of self-models. In W. Mack & G. Reuter (Eds.) *Social Roots of Self-Consciousness. Psychological and Philosophical Contributions* (pp. 161-186). Berlin, GER: Akademie 2009.
- Newen, A. & Vogeley, K. (2003). Self-representation: searching for a neural signature of self-consciousness. *Consciousness & Cognition*, 12 (4), 529-543. [10.1016/S1053-8100\(03\)00080-1](https://doi.org/10.1016/S1053-8100(03)00080-1)
- Noë, A. (2005). *Action in Perception*. Cambridge, MA: MIT Press.
- O'Regan, J. K. & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24 (5), 939-1031.
- Quadt, L. (2015). Multiplicity Needs Coherence – Towards a Unifying Framework for Social Understanding. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Siegel, S. (2012). Cognitive Penetrability and Perceptual Justification. *Nous*, 46 (2), 201-222.
- Vetter, P. & Newen, A. (2014). Varieties of cognitive penetration in visual perception. *Consciousness and Cognition*, 27, 62-75.
- Welpinghus, A. & Newen, A. (2012). Emotion und Kultur. Wie individuieren wir Emotionen und welche Rolle spielen kulturelle Faktoren dabei? *Zeitschrift für philosophische Forschung*, 66 (3), 367-392.
- Zinck, A. & Newen, A. (2008). Classifying emotion: a developmental account. *Synthese*, 161, 1-25.

Concept Pluralism, Direct Perception, and the Fragility of Presence

Alva Noë

This paper has three main aims. First, I criticize intellectualism in the philosophy of mind and I outline an alternative to intellectualism that I call Concept Pluralism. Second, I seek to unify the sensorimotor or enactive approach to perception and perceptual consciousness developed in O'Regan & Noë (2001) and Noë (2004, 2012), with an account of understanding concepts. The proposal here—that concepts and sensorimotor skills are species of a common genus, that they are kinds of *skills of access*—is meant to offer an extension of the earlier account of perception. Finally, I describe a phenomenon—fragility—that has been poorly understood, but whose correct analysis is critical for progress in the theory of mind (both perception and cognition).

Keywords

Actionism | Concept pluralism | Concepts | Consciousness | Enactive account | Evans | Fragility | Frege | Intellectualism | Kant | Perception | Plato | Presence | Sensorimotor account | The intellectualist insight | The intellectualist thesis | Understanding | Wittgenstein

Author

Alva Noë

noe@berkeley.edu

University of California,
Berkeley, CA, U.S.A.

Commentator

Miriam Kyselo

miriam.kyselo@gmail.com

Vrije Universiteit
Amsterdam, Netherlands

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

The present study takes its starting point from the enactive or sensorimotor, or, as I now prefer to call it, the actionist approach to perception and perceptual consciousness (O'Regan & Noë 2001; Noë 2004, 2012). Actionism is the thesis that perception is the activity of exploring the environment making use of knowledge of sensorimotor contingencies. Sensorimotor contingencies are understood to be patterns of dependence of sensory change on movement. The proposal, then, is that we make use of this knowledge of the way our own movement gives rise to sensory change to explore the world. This knowledge-based or skilful activity *is* perceiving.

We characterized the relevant kind of knowledge as *knowledge* precisely in order to mark the continuity between perception and “higher”, more intellectual kinds of cognition such as thought and planning (O'Regan & Noë 2001). At the same time, we were quick to characterize the relevant forms of knowledge as practical, non-propositional, as implicit, or as “skill”, precisely in order to avoid over-intellectualizing perception.

In *Action in Perception* (Noë 2004, Ch. 6), I defended the view that perception requires the mastery and exercise of concepts. In doing so, I took myself to be lowering the bar on what it is to have a concept, rather than raising the bar on

what it is to be a perceiver. It was always my view that the resulting account was one in which understanding (mastery and use of concepts, including sensorimotor skills) and perception (exploration of the environment drawing on a variety of skills, including concepts, as conventionally understood, and also sensorimotor skills) worked together in human and animal mental life. As I put it later, “understanding” and “perception” arrive at the party together (Noë 2012).

Although actionism places great emphasis in the importance of movement, action, and the body for the theory of perception, on the claim that perceiving is an activity, and on the proposition that perception is not a representation-building activity, it was never the intention of the view to deny the critical role of understanding and knowledge. The point, rather, was to offer a unified account of perception, consciousness, thought, and action. But the details were not entirely worked out. Knowledge, skill, ability, and understanding were not carefully defined, and the precise relation between the account of perception and that of conceptual understanding was not spelled out in detail. I try to rectify that here.

My basic strategy in this paper is as follows. In part I, I offer an extended discussion of what I call intellectualism. I define the view, criticize it, and show how even critics of the view tend to share many of its presuppositions. In part II, I try to offer an alternative to intellectualism, namely concept pluralism, which builds upon the actionist conception of concepts as “skills of access”. Concepts, I propose, should be thought of as techniques for enabling access to what there is. In this way—the details will become clear later on—I offer a way of thinking about concepts that is unified with the basic elements of the earlier theory of perception.

One caveat: I don’t take up the issue of animal experience and cognition in this paper, even though it is directly relevant to the topic.

I

2 Modes of understanding

Kant (1791) said that concepts are predicates of possible judgement. That’s what concepts are.

They are creatures of judgement. He also believed that concepts play a basic role in cognition. They organize the data of sense. Without concepts, sensory experience would be empty sensation; without sensory influx, there’d be nothing for concepts to organize. For Kant, judgement gives the basic form of experience (*Erfahrung*).

Frege (1891) said that concepts are functions from objects to truth-values. In this he appeared to break with Kant. Concepts have nothing to do with judgement or with our cognitive organization. They are before all that. This is in tune with Frege’s well-known anti-psychologism, according to which grasping, understanding, judging, and communicating are of no relevance to logic or ontology.¹ But Frege doesn’t actually sever the link between concepts and judgement; he only frames it differently. Concepts figure in what is judged; they belong to *judgeable content*. So Frege preserves Kant’s link to judgement, but in a de-psychologized version.²

Frege’s anti-psychologism gets him into trouble.³ The fact that concepts are not themselves psychological, in the sense of being ideas or associations or feelings, doesn’t mean that they are not tied to understanding or judgement, for nothing forces us to think of understanding and judgement as psychological in that sense. At the same time, the claim that concepts are “third-realm” entities gives little substance to the idea that they are, in the relevant sense, objective. Finally, if concepts are some sort of occult abstracta, then it isn’t at all clear how we can grasp them. And surely, whatever concepts are, it is the case that we can grasp them.

I’ll return to this set of issues later. But for now let us agree that for both Kant and Frege, concepts are tied to judgement, where this means something like: they are tied to categorizing, to explicit reasoning, to subsuming objects under concepts. Each of these thinkers offers an account of concepts, or of the under-

1 See, for example, Frege’s “Thoughts”, (1918–1919).

2 Not that I mean to suggest that it is right to think of Kant as actually offering a psychological account. But it might look this way from Frege’s perspective.

3 As both Dummett (1973) and Baker & Hacker (1984) have noticed.

standing of concepts, in what I'll call the mode of judgement. According to Kant and Frege, grasp or understanding of concepts finds its natural, true expression in judgement.

This paper takes its start from the observation that there would appear to be other modes of conceptual activity, other ways for understanding (for concepts) to find expression in our lives. At least on the face it, judgement would not seem to be the only mode of conceptual understanding.

Take, for example, *perceptual understanding*, or what we might call *understanding concepts in the perceptual mode*. Consider reading. It is difficult to tell, looking at the entrance to the Taj Mahal, which bits of squiggle are mere ornament, and which are writing in Classical Arabic. You can have this experience, it is available to you, only if you are not fluent in Classical Arabic, or in this style of Arabic script. This marks the spot of the basic phenomenon: there would seem to be a mode of understanding that is perceptual in nature. It is impossible, as a psychological matter, to see meaningful text as a mere squiggle. For the one who knows, for the one who can, meaningful words just show up.

Compare this with the case of a scholar studying Renaissance paintings in which writing is shown embroidered into the robes of magi and other fabulous figures. Are these scripts in a familiar language, or could they be marks from a forgotten one? Or are they *pseudo*-scripts? How do you decide? A keen problem and one that affords opportunity, for it demands reasoning, explicit categorization, and judgement.⁴

But nothing like that seems to be going on when you are reading. And the point is general: it operates at the level of our everyday seeing. It is difficult, maybe even impossible—psychologically speaking—to see familiar kinds of things around us as mere things. We always see them as this or that.

I don't mean that when we see, we *represent* the things we really see around us as this or that, by bringing them under the relevant con-

cepts, by categorizing them, as it were, in judgement. The point rather is that the things we see, the things around us, are familiar, known, comprehended, understood, and recognized, from the very outset. Concepts are geared in before we are even in a position to ask what something is or to make a judgement about it.⁵

So we have here a distinct way in which concepts, or the understanding, can be put to use outside the setting of judgement. Specifically, as I've said, this is an example of the deployment of concepts in the perceptual mode or, more simply, perceptual understanding.

Note, in saying perception is a non-judgemental mode of understanding, I don't mean to deny that there might be an interdependence between the judgemental and the perceptual modes. Maybe only one who can judge can perceive and precisely because perception enables judgement. And maybe it is only of one who can have perceptual experience that we could ever say that he or she is in a position to judge about anything.⁶ My point is that, on the face of it, judging is one thing, and perceiving another, and yet they are both ways of exercising the understanding.

There are other modes, as well.

Concepts also get deployed in what I call the *active mode*; understanding, that is, can find expression, immediately, in what we do. There is such a thing as *practical* understanding. And what makes the relevant understanding practical is not that it is an exercise in judgement on, as it happens, practical matters. What makes it practical, in my view, is that it is the gearing in or putting to work of one's understanding in the absence of any call for, or even space for, reflection or judgement.

The dog walker's knowledge of dogs, for example, is put to work in the way he or she adopts a gait that suits the dog and encourages or permits it to accomplish its sniffy, doggy business; and so also in the way the owner spontaneously shortens the leash as another dog approaches; it is exhibited, even, we might say, in

⁵ As Heidegger (1927) would have put it, the things we encounter are *always already* familiar.

⁶ I return to this issue of the unity of concepts in section 6 below.

⁴ For a discussion of this fascinating topic, see A. Nagel (2011).

the cool she keeps when the two dogs begin barking and straining at their leashes. Without a word, in the absence of deliberation, or explicit thought, the owner knowingly engages the nature of dogs.⁷

And there may be still other kinds of understanding, other *styles* of conceptuality. For example, there is also perhaps what we could call the emotional mode, or maybe it would be better to say the personal, or even *interpersonal* mode. Tears, feeling, injury, but also posture, standing distance to others, navigating in a social environment, can all show a highly refined attunement to situation, relationship, status, goals, tasks, and so on. It takes understanding to do all this, even though we rarely try to make this understanding explicit and even though, very probably, we cannot do this, even in ideal circumstances. Let us say that in this kind of responsive engagement with our social worlds we display understanding.⁸

To summarize: there is a case to be made for the existence of at least three, maybe four, distinct modes of understanding. There is the judgemental mode, the perceptual mode, and the active mode, and perhaps also the personal mode.

3 Intellectualism vs. the intellectualist insight

I have proposed that there are at least three or four distinct modes of understanding. I now turn to the familiar thought that among these varieties of expression of conceptual understanding, only one—the judgemental mode—is genuine. The other modes, according to this idea—that is, the perceptual, the active, the personal

—are expressive of understanding only derivatively, thanks to the fact that they are guided or controlled, from outside as it were, by true understanding in the judgemental mode.

I will call this view intellectualism. Intellectualism, as I am defining it, is the view that one modality of conceptual expression is basic, namely, the judgemental, and that the others are domains where understanding finds expression only derivatively.⁹

Plato and Descartes seemed to have believed something like this. For them, a mere sensation rises to the level of perception, and a mere movement to the level of action, only if it is subject to guidance by reason. The soul is divided against itself and it achieves integration only when it is controlled in the right way from above.

Intellectualism is probably the establishment view in cognitive science. When you see the Pole Star, for example, as [Fodor & Pylyshyn \(1981\)](#) insist, you represent whatever it is that you really see—a pattern of irradiation of the retina, perhaps—as the Pole Star. To suppose otherwise is to suppose that vision could be, as [Gibson \(1986\)](#) had claimed, a *direct pick up* of what there is around us. But Pole Starhood, like the third dimension, is not something that gets projected onto the retina. The whatness of things, their nature, no less than the third-dimension itself, are not, strictly speaking, visible. We need judgement, the application of concepts (in this case perhaps automatic and implicit) in the building-up of mental representations, to get something like the world into our experience.¹⁰

Jason Stanley, in a series of writings ([Stanley & Williamson 2001](#); [Stanley 2011](#); [Stanley & Krakauer 2013](#)), defends what I am calling intellectualism. You perform a skilful ac-

⁷ This example is from [Stephen Mulhall \(1986\)](#).

⁸ With this last example we move beyond description to the suggestion of an argument. The thought is that the relevant forms of understanding couldn't be underwritten by judgement, since we are not able, as a general rule, to frame the needed judgements. Indeed, something like this line of thought is already suggested in the way I've sketched the perceptual and active modes above. Recall the celebrated case of [Oliver Sacks \(1970\)](#): a man can't recognize the item before him as a glove; his powers of judgement are fine—he describes what he sees as a self-enclosed piece of fabric with five outpouchings—and he knows what a glove is. The case is illustrative because it brings out that it is less the fact that he can't recognize the glove, and more the very fact that he needs to think about it all, that brings home the thought that in our normal life there is no room for that sort of deliberation.

⁹ Intellectualism can be defined differently. For a variety of approaches to problems in this vicinity, see [Bengson & Moffett \(2011\)](#).

¹⁰ This was [David Marr's \(1982\)](#) view. The content of visual experience is given in a 2.5D sketch, that is, in a depiction of what is given in the projection of the world onto the retina. It is only in so far as vision yields *knowledge* that it goes beyond what is given in this intermediate-level representation and gives rise to a fully conceptual 3D model. But for Marr, and for his recent advocates ([Prinz 2013](#)), although we live in the world of the 3D sketch, our experience is confined to the intermediate-level representation. And crucially, for these thinkers, you don't need concepts or understanding at the intermediate level. You just need optics.

tion, according to Stanley (2011), only when your action flows from your knowledge of true propositions. He elaborates:

[t]here are all sorts of automatic mechanisms that operate in a genuine sense sub-personally. The human (and animal) capacity for skilled action is based upon these mechanisms. What makes an action an exercise of skill, rather than mere reflex, is the fact that it is guided by the intellectual apprehension of truths. (Stanley 2011, p. 174)

Is intellectualism right? Should we be intellectualists?

It is important that we notice, right away, that intellectualism is right about something. It does justice to the fact that there is understanding, and there is conceptuality, at work wherever we think and perceive and act and talk, as we have been considering. Conceptuality, understanding, and knowledge pervade not only the mental, but our lives and our being. Certainly, it is in evidence wherever we can speak of agency. Stanley insists (in the quotation above) that we can only speak of *skilful* action where there is understanding at work. He perhaps ought to have said that we can only speak of action at all, as opposed to mere reflex, or mere movement, where there is also understanding.

The question I would like us to consider is this: do we need intellectualism to secure this undoubted intellectualist insight, as I will dub the recognition of the pervasiveness of understanding in our perceptual, active, as well as emotional lives? It's crucial that we notice the distance between the insight and the thesis. It's one thing to say that there is understanding at work in perception and action, and another to think that what makes this true is that perception and action are grounded on acts of judgement. Do we need to think that what guarantees and secures the involvement of understanding is the fact that our seeings, doings, and feelings are guided by judgements?

There are, right off the bat, two obvious grounds for suspicion regarding the intellectual-

ist thesis. For one thing, intellectualism at least threatens to obscure the differences to which I have been directing our attention among what at least appear to be authentically distinct ways of exercising one's knowledge and understanding. And so, it seems, it gets things wrong. Seeing and acting and dynamically reacting, most of the time at least, don't look or feel anything like bringing objects under concepts in judgement.

For another, intellectualism smacks of the arbitrary. Couldn't we maintain that perception is the basic form of understanding and that judgement, even in cases of pure reasoning and mathematics, rests on a kind of perceptual insight? Or that it is understanding in the active mode that is truly basic? Judgement itself depends on the mastery and exercise of conceptual capacities which are in the first instance practical. You need to know how to use concepts, after all, in order to use them in judgement.

In any case, let us ask again: are there reasons to endorse intellectualism? Why think that judgement is the primary and singular authentic modality of real understanding? Why be an intellectualist?

4 Troubles with intellectualism

Stanley's writings (Stanley & Williamson 2001; Stanley 2011; Stanley & Krakauer 2013) on the topic are suggestive. However, he seems to mistake evidence in favour of the insight (that understanding is present in perception and action, as well as in the setting of explicit deliberative thought) with support for intellectualism itself (for the view that judgement governs action and perception). And, on top of that, he may commit the fallacy of conceiving the whole genus on the model of one of its species; like thinking that every dog is a cat because, well, they are mammals, or that seeing is a way of touching because, after all, they are both forms of perception. In this case it is the fallacy of thinking that *knowing how* must be a form of *knowing that* because, after all, it is form of knowledge.

Let's turn to this last point first, briefly. Stanley (2011) notices that we use "to know" both for propositional knowledge and also for

practical knowledge (know-how). Contrary to what he suggests, however, there are cognate languages where this is not the case. For example, we don't express knowing how in German using the same verb that we use to express propositional knowledge (Stanley 2011, pp. 36-37). We use *können*, which means *can*; we don't use *wissen* (as in *wissen wie*).

But in any case, the more important point is, *so what?* How dispositive are facts like this supposed to be? It is common ground, I would say, that know-how is a form of knowledge, an achievement of understanding. The question is whether it is a form of knowledge of the same type as propositional knowledge, the sort of knowledge that gets expressed in judgement. Crucially, all the evidence in the world that it is a form of knowledge doesn't add up to evidence that it is propositional knowledge.

Now, as a matter of fact, we know that knowing how to do something is *not* merely knowing that a proposition is true, for any proposition you might care to think up. For knowing how to do something implies that you have the ability to do it (and vice versa), whereas the corresponding propositional knowledge has no such practical entailments.

Stanley would deny this (Stanley & Williamson 2001; Stanley 2011). You can know how to perform a stunt but be unable to perform it (because you've been injured, say); so, he claims, possession of know-how cannot be equivalent to possession of an actual ability. But this is unpersuasive. Of course it is true that you can know how to do something even though you are unable to do it. But this is because your being unable to do it is not, in the relevant sense, evidence that you can't do it! Consider: you can't swim if there's no water, even though you can swim. You can swim but you can't swim. Far from showing that know-how and ability part ways, this sort of consideration reminds us that they move along the same rails.

So knowing how to do something isn't possession of propositional knowledge: it doesn't consist in being in a position to make certain judgements. This is a point that Stanley and Williamson accept, if only implicitly, for they provide a different analysis of the cases precisely

to account for the critical link to action in the case of know-how. Knowing how to do something, on their view, consists in grasping a true proposition, yes, but it consists in grasping it in a distinctively and irreducibly practical way (making use of practical modes of presentation).¹¹

Again, it is worth noticing that to deny, as I do, that knowing how to do something consists in knowing the truth of a proposition, is not to deny that, as a matter of fact, knowing how to do something may put you in a position to make certain judgements, or may require you to appreciate the truth of certain propositions.

This brings us to the first point above: the confusion of evidence for the insight with evidence for the thesis. I am assuming that know-how, like propositional knowledge, is a form of knowledge. This common ground is already secured by the insight: our understanding, our knowledge of concepts, is put to use in both cases. So we can readily agree with Snowdon (2004), cited approvingly by Stanley (Stanley & Williamson 2001), that knowing how and knowing that go together—that where you have one, you have the other. In general, as Snowdon observes, if you know how to do something—say, how to get home from here—then you'll know that all sorts of things are true, such as, for example, *that* you need to turn left here, that you aren't already home, etc. And vice versa. Knowing how and knowing that, in this sense, commingle and cooperate. These considerations are adduced by Stanley, and by Snowdon, I think, to suggest that Ryle was mistaken in believing that the propositional and the practical are disjoint and disconnected (1949); in fact they operate together and in support of each other. This is an important point and one I endorse. And this is exactly what one should expect given the intellectualist insight. After all, understanding operates in both spheres: the practical and the judgemental or propositional. Crucially, however, the fact that the practical and the propos-

¹¹ Stanley (2011) offers a different account from that developed in Stanley & Williamson (2001). The former is framed in terms of modal parameters governing the interpretation of the relevant sentences. Although he insists that know-how does not entail ability, he admits that attributions of know-how exhibit more or less the same sort of modality as ascriptions of dispositions and abilities.

itional mutually entail each other in this sort of way lends no support to the intellectualist idea that one of these, the propositional, is foundational in respect of the other; indeed, it weighs against that very idea. Why press on and insist on this thesis when, it would seem, the insight on its own is enough to capture the phenomenon at hand?

Stanley's motivations seem fairly clear. He wants to break with the idea that propositional knowledge is detached and, as he puts it, behaviourally inert. He wants to insist that it's wrongheaded to think that athletes and clowns and craftspeople are skilful zombies, whereas philosophers and mathematicians and physicists are *intellectual* workers whose actions exhibit authentic brain-power. It may be, even, that he thinks this is a point of political significance.

Intellectualism isn't necessary to secure any of this, however. The insight has already done that.

In fact, intellectualism, as Stanley develops it, threatens to distort the nature of the cognitive achievements that are put to work in our practical, perceptual, and personal engagements. This comes out in the discussion of skill. Stanley & Krakauer (2013) defend Aristotle's claim (from *Metaphysics* 1046b) that we can only speak of *skilful action*, as opposed to mere habit, or brute capacities, where we can speak of *rational control* of action, and also where we can speak of teaching, learning, practicing, getting better, or achieving expertise. They defend Aristotle's claim that it is a mark of skilfulness, that you can voluntarily choose to perform what you can do skilfully *badly*.

This last point seems unlikely. I can't choose not to understand what you say, or to see writing as mere squiggles, or words as composed of bits I need painstakingly to sound or spell out. A guitarist cannot choose to experience the instrument in his hands as strange or unfamiliar. At best, maybe, I can pretend I am unable to do these things.

Is this because talking and reading and playing guitar are not really skilful at all, that they are mere habits outside the range of rational control? Hardly! They're expressions of skilful competence, rational understanding and

knowledge if anything is. The mistake is to think that a performance is only rational if control is exerted in the mode of judgement, as if from outside. The understanding that is put to work in our talk and play, as in our thought, is native to these various styles of engagements themselves.

Stanley and Krakauer make a lot of the demand that skill depends on knowledge of facts. It's worth noticing, yet again, that insisting, as I do, that skilfulness does not consist in the *exercise of concepts in the judgemental mode* does not entail that there can be skilfulness in the absence of the ability to exercise them in that mode. It may be, as a matter of fact—this is related to the Snowdon point above—that only someone who is sensitive to all sorts of facts, for example, about how something is done, will in fact know how to do it. This doesn't show that knowing how is a kind of knowledge of the facts. It shows rather that our distinct conceptual capacities may be interdependent.

Stanley and Krakauer try to draw a line between true skills, which are, in their sense, governed by rationality, and others—for example perceptual and linguistic skills—that are too basic, or too simple to qualify as skills in the fuller rational sense.¹²

One problem with this suggestion is that it is not so easy to draw a sharp line between skills and supposedly brute abilities. Take colour vision, for example, which is innate in humans. Despite this, it turns out that children find it very difficult to recognize and discriminate colours long after they've mastered the names of familiar objects, people, games, etc. As Akins (unpublished manuscript) has argued, this is probably because colours are not simple, as our phenomenology, or rather, our conventional wisdom about our phenomenology, leads us erroneously to believe. *Getting* blue or yellow or red is to develop a sensitivity to suites of constancies and variations—to ecological variation in what I have called *colour-critical condi-*

¹² Stanley & Krakauer (2013, p. 5) write: "[b]ut at some point, all such knowledge will rest on knowledge of basic actions, such as grasping an object or lifting one's arm. These activities are not skills; they are not acquired by or improved upon by training in adult life. Their manifestation is nevertheless under our voluntary control."

tions—that takes time and learning, and allows for criticism and reflection. Is colour vision basic? Or is it skilful? It may be both.

This is not a special case. Because seeing is saturated with understanding, it is very hard to find features of our ability that are not modulated by knowledge and context. Granted, the ability to discriminate line-gratings of different densities is fixed, at its limit, by the resolving powers of the eyes; yet our discriminations are likely to be sensitive to task and motivation, to attention and distraction—that is, very broadly, to our engagement with the meaningful world. So where does skill stop and brute ability begin? I am skeptical that learnability, teachability, or rational control provide an interesting or valuable demarcation. The most basic reason for this is that perceiving is never merely registration. It is a matter of knowledgable access (Noë 2004, 2012).

There is a second important issue as well. Consider language. Linguistic misunderstanding doesn't stop language in its tracks, ejecting you and sending you back to the grammar, written, as it were in advance, by those responsible for setting up the language. Rather, coping with misunderstanding—dealing with not getting how someone is using words, or how we should use them, or with not knowing how to use them—is one of language's familiar settings. We adjudicate and teach and learn and improve and criticize and define and formalize and evaluate *within* language, not from outside it. Language, contrary to the claims of Chomskyan linguistics, is not a rule-governed activity. It is a rule-using activity. And we make up the rules as we need them and for our own purposes. This may be controversial. But here's why I insist on it: according to the logician's or the linguist's picture of language, first you assign values to primitives, then you set up rules governing the construction of well-formed formulas. If you think of language this way, then it looks like you need judgement—the application of rules to cases—to secure the meaningfulness of what would otherwise be mere marks and noises. But we don't need judgement—we don't need understanding in the judgemental mode—to secure meaning. We don't need guidance from the outside.

The opposition between habit and skill is a false one; and it is a mistake to think that what marks the opposition is that habit is below or before understanding whereas skill is the deliberate exercise of understanding.

5 Troubles with anti-intellectualism

Some critics of intellectualism argue that perception cannot be conceptual, because if perception were conceptual, then perception would be a form of judgement. But the idea that perception is judgement over-intellectualizes perception.¹³

This is how I understand Gareth Evan's (1982) argument in connection with the Müller-Lyer illusion. You can experience the two lines in the Müller-Lyer illusion as different in length, even when you know, and so have not the even the weakest inclination to deny, that the lines are the same in length. The visual experience is one thing, and judgement another; hence experience is not conceptual.

Now, this is an example of an apparent disagreement between what you know to be the case (judgement) and how things look (experience). Things look precisely the way you know they are not. Experience and the judgement are in conflict. This shows, I would have thought, that experience, and the corresponding content, share the same kind of content. The fact that they are in apparent conflict shows that they are not somehow incommensurable. So if the one is conceptual, then so is the other.

But more important, for our discussion here, is that Evans seems to assume that concepts can only be in play if they are applied in judgement. Since experience is not judgement, there is no way for concepts to gear in. But that's to accept the basic claim of the intellectualist—judgement is the only way for concepts to get into the act—not to challenge it.

So Evans' argument against the idea that perceptual experience is conceptual—what we can think of as Evans's anti-intellectualism—actually takes what I am calling intellectualism

¹³ See Noë (2004, Ch. 6) for detailed engagement with the issue of the conceptuality of perception and the relation between my own position and that of John McDowell.

for granted. It takes for granted that there is only one genuine and legitimate mode of exercise of conceptual understanding, namely the judgemental.

Hubert Dreyfus (e.g., 2013) is responsible for a widely-influential criticism of intellectualism that is *crypto-intellectualist* in just this way.

Reasons, principles, and explicit knowledge guide perception and activity, according to Dreyfus, but only in the case of the novice. The expert, in contrast, is one who is engaged, in the flow. The expert, having mastered the rules and the concepts, has no further use for them. The expert is able to respond to the solicitations of situation and environment with no need for conscious thought or deliberate judgement.

A favourite example is that of the lightning chess player. There is literally no time, claims Dreyfus, for the chess player to analyse the situation and decide how to move. Moves are made in a flash. To suppose that the move is guided by reasons or judgement is to fall prey to a *myth of the mental*, according to which a mind-faculty, a faculty of judgement, say, accompanies our doings and is responsible for them being expressive of competence, intelligence, and understanding. For Dreyfus this idea is a dead giveaway of a distinct type of intellectualist psychologism. Yes, Dreyfus grants, if you ask the expert afterwards, why he or she made this move and not that one, he can give you a reason. But we have no more ground to suppose the reason was in operation before the player switched into the intellectual mode in response to the question than we do to suppose that the refrigerator light is always on because it is on whenever you open the fridge to look.

According to Dreyfus, understanding or reason operate only if there are explicit acts of rule-following, or judgement, that accompany, or even precede, every act. But why believe that? The baseball player doesn't need to be thinking about the rules for it to be the case that what he does is subject to them and is carried out, so to speak, in their light. The rules are there—in the form of umpires and rule books, and also dictionaries and courts of law, and earnest disagreement among participants—and we have access to them as need arises. The

fact that we can use them, and that we care about their correct use, is all that is needed for it to be the case that we act under their influence. The influence is not causal. It is normative.

Dreyfus goes further and insists that whether or not it is always legitimate to demand that *the phronesis*, as he calls the expert, invoking Aristotle, justifies his or her actions, it will not in general be possible for him or her to do so. You can't make explicit the myriad rules governing how we stand or react or explore or decide because, as a matter of fact, there are no such general rules. There is nothing to be made explicit. At best the chess master is likely to point to the situation on the board and exclaim, *look! This situation requires this move!*

But why is not this exactly the kind of reply that is required? Recall Wittgenstein's (1953, §88) example of "Stand over there!" This can be a perfectly precise command, as exact as rationality can require, even when it is not the case that one can specify, to the millimetre, say, where it is one is supposed to stand. For certain purposes, in certain contexts, one may need more precision. But in other contexts the demand for precision on the order of millimetres would be unreasonable. And so my thought here is that it is to set too high a standard on what it would be to have a reason for acting to demand that one can frame it independently of the situation one is in. It is precisely an over-intellectualized conception of what it would be to have a reason, or to make use of a rule, to suppose that rules and reasons need to be context-free and situation-independent, known in advance and applied, as it were, from outside one's engaged play¹⁴—just as it would be to over-intellectualize the intellect in general to suppose that concepts only gear in in the setting of judgement.

Here's the point: the use of rules themselves—which for Dreyfus is the hallmark of the detached attitude of the intellect—is itself an activity that admits of mastery and expertise and so also flow. And so we cannot insist that rule-use marks the boundary between engagement and detachment.

¹⁴ See McDowell (1994). His discussion of demonstrative senses and demonstrative concepts aims at just this point.

But once we allow that rules are used, and reasons proffered, from the standpoint of our engagement—from the inside—, then we need not fear that we have committed ourselves to an over-intellectualized conception of what it is to be engaged, just because we allow that we understand and can reflect on what we are doing.

Notice again that Dreyfus's picture—a picture he may take over from Heidegger (1927) and Merleau-Ponty (1945)—only counts as evidence against the idea that concepts and reasons and rules gear into perception and skilled action if we suppose that the intellectualist is right, that there is only one way for understanding to get into the act—namely, in the form of explicit deliberate judgement.

And notice that this way of rejecting intellectualism—on the part of Dreyfus, and other existential phenomenologists, and perhaps also Evans—pays a high price. For it must reject the idea that understanding and reason have any place at all outside the range of explicit deliberative reason, and so it has to give up the intellectualist's insight, namely that in our engaged, perceptual, and active lives, even when we are experts, even when we are skilled, our performance gives expression to knowledge, intelligence, and understanding. By accepting the intellectualist thesis that judgement alone is the only true way for concepts to gear in, Dreyfus and co. feel they are compelled to reject the idea that our lives as a whole, beyond the confines of deliberate exercise of reason and understanding, can be, or are, at one with our intellects.

What existential phenomenology may find difficult to appreciate—at least in Dreyfus's version of the position—is that conflict, disagreement, and disturbance of flow are themselves business-as-usual; they are normal moments in the way that even the expert carries on. We saw this in the language case. Expertise is not immunity; if anything, it is an evolved opportunity for new forms of vulnerability. Engagement is, as I shall put it, always manifestly *fragile*. That is, the liability to slip up, to get things wrong, is a built into the nature of the undertaking—of *any* undertaking. To go wrong is not, as a general rule, to stop playing the game—it is not the game's abeyance—it is rather a moment in

the development of play. But let's go back to language. We don't stop communicating when we fail to understand each other. At least that is not usually the case. Misunderstanding is an opportunity for more communication. Clarifying, reformulating, trying again, like criticism, are things we use language to do. The fragility is intrinsic and manifest. It doesn't mark out the game's limits. It marks one of its modalities.

I stated earlier that understanding in the active and perceptual modes leaves no room for the application of understanding in the judgemental mode. I suggested this was a reason for thinking that judgement can't be operating behind the scenes when we perceive and act. But we can amend this now in light of our consideration of fragility. It is internal to the very character of our perceptual and active involvements that they are liable, not so much to breakdown, in Dreyfus's sense, as to error, confusion, and other stutter-steps that require precisely that one now *think* about what one is seeing and what one is doing. Judgement and thought can, in this sense, live cheek-by-jowl with perception and action without, therefore, getting in their way.

In any case, Dreyfus's criticism of intellectualism fails. But it does so precisely because he fails to break with the over-intellectualized conception of the intellect at the heart of intellectualism. Dreyfus's anti-intellectualism fails because intellectualism fails. It is, in reality, a species of intellectualism. Neither Dreyfus, nor his would-be opponent, can do justice to the ways in which understanding operates outside the narrow domain of explicit reasoning. Both sides fail to accommodate the phenomenon of fragility.

II

6 Concept pluralism: A genuine alternative to intellectualism

So let us now turn our attention to the prospects for framing a true alternative to intellectualism. What would such an alternative look like?

A genuine alternative to intellectualism will be pluralist in that it will reckon that there

are different legitimate and non-derivative modes of understanding, and so it will hold fast to the intellectualist's insight that understanding is in play everywhere in our lives even as it rejects the intellectualist thesis.

One resource for such a pluralism is [Wittgenstein \(1953\)](#). Wittgenstein proposed that a concept is a technique, and that understanding, therefore, is a form of mastery, akin to an ability. An important fact about abilities is that they can be exercised in a multiplicity of ways. I can exercise my understanding of what a house is by building one, looking at one, painting one, living in one, talking about one, or buying one. So, from this standpoint, there is nothing more surprising about the fact that my knowledge can find expression in what I do, as well as in my knowledge of a proposition, than there is in the fact that my ability to read gets exercised both when I read a novel and also when I blush at the words on the bathroom wall.

This idea also helps us explain the unity of understanding. If concepts can be applied in walking the dog as well as in writing a treatise about dogs, what is the connection between these two self-standing and non-derivative modes of exercise of something that, surely, is a single conceptual capacity: an understanding of the concept *dog*? What gives unity to this understanding?

The idea that understanding a concept is mastery of a technique, a mastery that has multiple, distinct, context-sensitive ways of finding expression, helps here. One way to express understanding of *dog* is to talk and write about dogs. Another way is to be able to spot dogs on the basis of their appearance. Still another is to work or play comfortably with dogs. And the list goes on and on. We put our singular understanding of what dogs are to work in these different ways, and the understanding consists in the ability to do (more or less) all of that.

We are now in a position to appreciate that the claim that perception and action are, with judgement, non-derivative, original modes of understanding does not entail that these modes are independent of each other. The idea that the unity of a concept is a matter of unity-in-ability helps bring this out. The fact that

perception isn't beholden to judgement for its conceptuality doesn't mean that there could be perception in the absence of capacities for judgement. After all, typically, you can't be said to know a concept if you can't apply it in normal perceptual settings. Can you know what a tomato is if you are incapable of any active or perceptual engagement with tomatoes?

But we should also be careful. In so far as our concepts have unproblematic unity, then, on this Wittgensteinian view, this is because they are exercises of common abilities—abilities which are, of their nature, such as to admit a genuine multiplicity of expressions. But the unity of our concepts is not something that we can always take for granted.

Is there *one* concept of dog, or several, brought to life in different situations and sub-cultures at different times, for different purposes? Is there unity or just fragmentation? Is this a shared understanding? These are important questions, not for philosophy, particularly, but for culture. Look at the changes that have taken place in our thinking about *matter* over the last few hundred years. Or, to give a different kind of example, about *gender*. We have no choice but to work it out as we go along.

And crucially, there is no standpoint outside our thinking, talking, writing, persuading, imposing, regulating, prescribing and also describing, from which these questions can be adjudicated. This doesn't make the existence of dogs a matter of social construction. (Of course, dogs are, literally, bred and so constructed by us.) No, surely dogs have a mind-independent nature. But it does mean that it is hard and creative and unending work to bring that reality into focus in our shared thought, talk, perception, and activity.

There is no standpoint outside our thoughtful practices from which to ask after our own concepts. For our concepts are our own tools and techniques. This is where Frege went wrong. He seems to have thought that the only way to achieve *objectivity*—that is, sharability, articulability, and lawfulness—was by supposing concepts were out there, indifferent to how we grasp or understand them. In fact, they supervene on our grasping, negotiating, communicat-

ive activity. Frege made no allowance for fragility.

7 Concepts are skills of access

But can we say more than just that concepts are abilities? Abilities to do what? Well, we've already said: to talk and see and use and judge, and so on.

But I think we can do better. To do so, I draw on the actionist approach to perception developed in earlier work (Noë 2004, 2012). To begin to organise an answer, consider two familiar facts about visual perception. The first is that, as Euclid noticed, when a solid opaque object is seen, it is never seen in its entirety at once. Things always have hidden parts. The second is that the visible world is cluttered with all manner of stuff. Things get in the way, the view is interrupted, occlusion is the norm.

And yet, despite these striking limitations, we don't experience the world as cut off from us, inaccessible to vision, blocked from perception. The partial, fragmentary, and perspective-bound character of our visual access to the world is not a limit on what we see, a marking off of our liability to blindness; it is, rather, the very manner of our seeing. This is fragility again.

Not seeing through the solid and opaque, as if it were transparent, is not a perceptual failing but rather an accomplishment. And relatedly: we belong to the cluttered environment ourselves. We are not confined to what is projected to a point. We explore. And it is that exploring, that doing, that is the seeing. The seeing is not the occurrence of a picture or representation in the head; it is, rather, the securing of comprehending access, thanks to our possession of a specific repertoire of skills, to what there is. The generic modality of the way the world shows up in perception is not *as represented*, but rather *as accessible* (as I argue in Noë 2012). This is why our inability to see things from all sides at once, or to experience a thing's colour in all possible lighting conditions at once, is no obstacle to the presence of whole objects and colours in our experience.

The immediate environment is present in visual perception, not because it projects to the eyes, but because the person, by means of the use of his or her eyes as well as other forms of movement and negotiation, has access to that environment. Presence is availability, and its modalities—visual as opposed to tactual, for example—are fixed by the things we need to do, the negotiations, to bring and keep what is there in reach. Wittgenstein, in the *Tractatus* (1921), said that the eye is a limit of the visual field. But this is wrong: the adjustments of the eye, the need to adjust the eye, difficulties in adjusting the eye, are given in the way we see. Wittgenstein's point, I suppose, was that the eye doesn't see itself seeing (unless you look in a mirror). But here's a different model: seeing is like what an outfielder does. To say that the eye is not *in* the visual field is a bit like saying that the body of the outfielder is not in the field of play. But in fact the eye and the head and the hand and the arm and the glove are all in the field of play. And what we call *fielding the play* is precisely a temporally extended transaction in that whole environment. And the basis of the environment's availability to this or that modality of exploration, beyond the fact that it is there, is our possession of the skills, abilities, and capacities to secure our access to it. The occluded portions of the things we see are there for us, present to us, thanks to our skilful ability to move and bring them into view. Perception is fragile.

John Campbell, writing in a related context (2002), has said that we shouldn't think of the brain as representing the world; we should think of it as making the adjustments that, as he puts it, keep the pane of glass between you and the world clean and clear, as if it were continuously vulnerable to becoming opaque.

My thought is that *we* (not our brains) need continuously to make adjustments to keep the world in view, and to maintain our access to the world around us.

But I add: the character of the world's presence itself is precisely a function not only of what there is, but of what we know how to do, and what we do, and what we must always of necessity stand ready to do, just in case, to pre-

serve our access. You need to squint and peer and adjust to see things far away; and this makes a difference to how those things show up.

This is one reason why it is a mistake to suppose that we think of the adjustments that belong to the ways we bring the world into focus as the brain's work. No, it is our work, even if most of it is low-level, unattended, and done automatically. For it is this work that gives experience the quality that it has.

The scene is present for us in the manner of a field of play. This is a fragile presence. Its presence is not given to us alone thanks to what might happen in our brains, thanks to neural events triggered by optical events. Its presence is achieved thanks to what we know how to do. The basis of our skilful access to the world is, precisely, our possession of *skills of access*.

And this, finally, is what I propose concepts are. They are skills of access, or rather, a species of such. They are not so much devices by which we make the world intelligible, as much as they are the techniques by which we secure our contact with the world, in whatever modality. From this point of view, concepts like *dog* and *matter* are of a piece with other skills of access such as the not-quite-articulable sensorimotor skills we skilfully deploy as we navigate the scene with our thinking bodies.

From this standpoint, it is worth emphasizing that there is no theoretically interesting cleavage between seeing and thinking (as already argued in Noë 2012). Seeing is thoughtful and thought is perceptual at least in so far as it is, like seeing, a skilful negotiation with what there is, as just another modality of our environment-involving transactions. Presence, after all, is always in a modality—that is, it is always dependant on our repertoire of skills. And it is always a matter of degree. The hidden portions of the things we see *show up for us*, as does the space behind our head, and even spaces further afield. We have access—skill-based, partial, perspective-bound, and fragmentary—to it all.

Perception and thought, from the actionist perspective, differ as sight and touch differ. They are different *styles* of access to the world around us.

8 We use concepts to take hold of things, not to represent them

Let us come back to the more particular line of investigation that has been our concern.

The intellectualist is quite right that in so far as seeing is expressive of understanding, this is because we bring concepts to bear in our seeing. But the intellectualist is mistaken in holding that this is because we categorize what we see, in the mode of judgement, by applying concepts. It is rather that we see *with* concepts. Concepts are techniques by which we take hold and secure access. Their job is not to represent what is there; their job is to enable what is there to be present to us. You can't see the laser-projector if you don't know what a laser-projector is. Your possession of the concept is a condition on the laser-projector's showing up for you. It is the ability that lets you encounter what is in fact there.

Back to the example of text: your grasp of the relevant concepts enables you to read (to see what is there). Not because it gives you the resources to interpret or decode (although it does give you that). But because knowledge lets what might otherwise be unseen come into view. Knowledge can also, correspondingly, disable us. Your reading knowledge, for example, can make it difficult or even impossible to see the squiggles, the “mere marks”, which are also always there whenever you read.

And so across the board: we don't apply concepts in judgement to what we see in order to represent things; our possession of the concepts is what enables us to make contact with them themselves. We see *with* our concepts. They are themselves techniques or means for handling what there is. Think of the concept in perception not as a category, or a representation, but a way of *directly picking up* what is there (to re-use and rehabilitate Gibson's 1986 idea).

And so also for the active modality. My understanding gets expressed in what I *do* and it gets expressed directly—for example, I exercise my knowledge of teacups in the *way* I handle this cup; I grasp the cup with my hands, and also with my understanding. My under-

standing gets put to work in the fact that I am able to do this, in the fact that I know how to do it.

Understanding, I would urge, is put to work, in these doings, *directly*. We don't need to suppose an action is skilful or knowledgeable or expressive of understanding only when it is guided, as it were from without, by propositional knowledge—as if the understanding couldn't inform our practical knowledge and our action directly.

And we are now finally in a position to understand why this is the case: for then we would be owed an account of how understanding is put to work in judgement. And here, we are just thrown back on what we can do to bring what is there for us into focus, to achieve its presence.

9 Conclusion: The significance of fragility

The world shows up for us in perception and thought, but it has a fragile presence. It shows up in very much in the same way that what a person means shows up for us when we are in conversation, to return to the language example. Misunderstanding, outright failure to understand, are always manifestly live possibilities. It isn't only solid opaque objects that fail to reveal themselves in their totality to the single glance. What we are given, always, is an opportunity or affordance for further effort, engagement, negotiation, and skilful transaction. The world is present to thought and perception not as a represented totality—an idea in our minds, a representation in our brains—but as the place in which we find ourselves, where we live, where we work. The world is a big place, and so there is a lot for us to do if we are to secure our footing on its slippery grounds. But a slippery ground is still a ground, and we need to secure our footing.

Presence—in thought and experience—is fragile, in other words. Philosophy has been strangely resistant to fragility. Fragility is not fallibility. The point about fragility is that it is manifest. An object's colour shows up for us as something with hidden aspects; it presents itself

to us as something that is always on the cusp of variation, always ready to change with the least alteration in our perspective or in the conditions of viewing. A colour, no less than a solid object, has hidden aspects. We don't experience these aspects as isolated atoms—as if we were confined to what the camera sees. What we see, what we experience, outstrips anything that can be understood in optical terms alone. For we see, we experience, and we also think about, a world that manifestly goes beyond what can be taken in a glance. Our skills—our understanding, to use the term that has organised so much of this discussion—gives us access to what there is.

That access is achieved, but not once and for all. It is not as though we consume the world in encountering it so that now we can make do with what is inside us. Access is a work in process. Presence is fragile, manifestly so; but it is robust.

Acknowledgements

I have presented this paper at Georg-August-Universität Göttingen, Ruprecht-Karls-Universität Heidelberg, the University of Iowa, the University of Pittsburgh, Yale University, and also in Riga at the Riga-Symposium on Cognition, Communication and Logic in May 2013, as well as at the 2014 Wittgenstein Symposium in Kirchberg am Wechsel. I am grateful to these audiences for their helpful comments and questions. For comments on the talk, or on the written paper itself, I would particularly like to thank Michael Beaton, Andy Clark, James Conant, Caitlin Dolan, Hubert Dreyfus, Sean Kelly, John W. Krakauer, Zachary C. Irving, Edouard Machery, Thomas Ricketts, Jason Stanley, David Suarez, and Martin Weichold.

References

- Akins, K. (unpublished manuscript). *Unpublished manuscript. Presented at Riga Symposium*. Riga, Latvia.
- Aristotle, (1924). *Metaphysics*. In W. D. Ross (Ed.) *Aristotle's metaphysics*. Oxford, UK: Clarendon Press.
- Baker, G. P. & Hacker, P. M. S. (1984). *Frege: Logical excavations*. Oxford, UK: Blackwell.
- Bengson, J. & Moffett, M. A. (Eds.) (2011). *Knowing how: Essays on knowledge, mind, and action*. Oxford, UK: Oxford University Press.
- Campbell, J. (2002). *Reference and consciousness*. Oxford, UK: Oxford University Press.
- Dreyfus, H. (2013). The myth of the pervasiveness of the mental. In J. K. Schear (Ed.) *Mind, reason, and being-in-the-world: The McDowell-Dreyfus debate*. London, UK: Routledge.
- Dummett, M. (1973). *Frege: Philosophy of language*. Cambridge, MA: Harvard University Press.
- Evans, G. (1982). *The Varieties of Reference*. Oxford, UK: Oxford University Press.
- Fodor, J. A. & Pylyshyn, Z. W. (1981). How direct is visual perception: Some reflections on Gibson's "ecological approach". *Cognition*, 9 (2), 139-196. [10.1016/0010-0277\(81\)90009-3](https://doi.org/10.1016/0010-0277(81)90009-3)
- Frege, G. (1891). Function and concept. *Collected papers on mathematics, logic and philosophy* (pp. 137-156). Oxford, UK: Blackwell.
- (1918). Thoughts. *Collected papers on mathematics, logic and philosophy* (pp. 351-372). Oxford, UK: Blackwell.
- Gibson, J. J. (1986). *The ecological approach to visual perception*. Princeton, NJ: Lawrence Erlbaum Associates.
- Heidegger, M. (1927). *Being and time*. New York, NY: SUNY Press.
- Kant, I. (1791). *Critique of pure reason*. London, UK: Macmillan.
- Marr, D. (1982). *Vision*. San Francisco, CA: W.H. Freeman.
- McDowell, J. (1994). *Mind and world*. Cambridge, MA: Harvard University Press.
- Merleau-Ponty, M. (1945). *Phenomenology of perception*. London, UK: Routledge.
- Mulhall, S. (1986). *Heidegger and being and time*. London, UK: Routledge.
- Nagel, A. (2011). Twenty-five notes on pseudoscript in Italian art. *Res: Anthropology and Aesthetics*, 59/60, 228-248.
- Noë, A. (2004). *Action in perception*. Cambridge, MA: MIT Press.
- (2012). *Varieties of presence*. Cambridge, MA: Harvard University Press.
- O'Regan, J. K. & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24 (5), 883-975. [10.1017/S0140525X01000115](https://doi.org/10.1017/S0140525X01000115)
- Prinz, J. (2013). *The conscious brain: How attention engenders experience*. New York, NY: Oxford University Press.
- Ryle, G. (1949). *The concept of mind*. London, UK: Hutchinson's University Library.
- Sacks, O. (1970). *The man who mistook his wife for a hat, and other clinical tales*.
- Snowdon, P. (2004). Knowing how and knowing that: A distinction reconsidered. *Proceedings of the Aristotelian Society*, 104 (1), 1-29. [10.1111/j.0066-7373.2004.00079.x](https://doi.org/10.1111/j.0066-7373.2004.00079.x)
- Stanley, J. (2011). *Knowing how*. New York, NY: Oxford University Press.
- Stanley, J. & Krakauer, J. W. (2013). Motor skill depends on knowledge of facts. *Frontiers in Human Neuroscience*, 7 (503). [10.3389/fnhum.2013.00503](https://doi.org/10.3389/fnhum.2013.00503)
- Stanley, J. & Williamson, T. (2001). Knowing how. *Journal of Philosophy*, 98 (8), 411-444.
- Wittgenstein, L. (1921). *Tractatus logico-philosophicus*. London, UK: Routledge.
- (1953). *Philosophical investigations*. Oxford, UK: Blackwell.

The Fragile Nature of the Social Mind

A Commentary on Alva Noë

Miriam Kyselo

In this paper I argue that while Noë's actionist approach offers an excellent elaboration of classical approaches to conceptual understanding, it risks underestimating the role of social interactions and relations. Noë's approach entails a form of body-based individualism according to which understanding is something the mind does all by itself. I propose that we adopt a stronger perspective on the role of sociality and consider the human mind in terms of socially enacted autonomy. On this view, the mind depends constitutively on engaging with and relating to others. As a consequence, conceptual understanding must be seen as a co-achievement. It is a fragile endeavour precisely because it depends not only on the individual but also on the continuous contribution of other subjects.

Keywords

Body-social problem | Enactive self | Fragility | Socially enacted autonomy | Socially extended mind

Commentator

Miriam Kyselo

miriam.kyselo@gmail.com

Vrije Universiteit
Amsterdam, Netherlands

Target Author

Alva Noë

noe@berkeley.edu

University of California
Berkeley, CA, U.S.A.

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

In the paper "Concept Pluralism, Direct Perception, and the Fragility of Presence" Alva Noë offers an exciting and dense insight into his philosophical thinking. Combining his classical work on the active nature of perception (Noë 2004) with his more recent inquiries into philosophical method, presence, the arts, and human nature in general, Noë now aims at a more thorough account of conceptual understanding (2012).

Noë's proposal must be seen in light of the paradigm shift in the philosophy of mind and cognition, from a cognitivist and representationalist view to a distributed or embodied per-

spective on the mind. It is one of the so-called "E-approaches" to the mind (enactive, extended, embodied and embedded) that transcend the classical view of the mind as being an isolated entity located in the brain that passively represents an outside and independently-given world (e.g., Shapiro 2011; Clark & Chalmers 1998; Noë 2004; Varela et al. 1993; Thompson 2007; Kyselo 2013). There are significant differences between these views (and they will be of relevance below), but generally speaking they all rest on the assumption that cognition is not in the head and instead requires bodily action and the environment. Noë uses these insights

from the E-approaches to expand on the disembodied and representationalist view underlying the intellectualist approach to concepts, and in this way, he provides a timely and innovative elaboration of conceptual understanding that is more encompassing than previous approaches.

I am sympathetic to Noë's approach. Methodologically speaking, he illustrates what he promotes as the right style of philosophical analysis, an inquiry into the so-called "third-realm" that remains "in-between—neither entirely objective nor merely subjective" (Noë 2012, p. 136) but open for "conversation or dialogue" (Noë 2012, p. 138). My comment should be considered an elaboration in the same vein.

I agree with Noë with regards to the more general project of questioning traditional conceptions in philosophy of mind by adopting an embodied and distributed perspective. That said, however, I think that there is a problem with his proposal. Even though it provides a great number of important insights, I think, third-realm fashion, Noë's proposal fails as a general theory of understanding. The reason for this is that in a crucial way his own epistemological pre-conception of mind is not yet fully separated from the paradigm that it seeks to overcome: while Noë acknowledges the role of the bodily and active individual, he accepts a dichotomy that is prevalent in the traditional paradigm, namely the split between the individual and the world of others. His approach inherits what I have called the *body-social* problem (Kyselo & Di Paolo 2013; Kyselo 2014). The body-social problem is the third in a series of dichotomies in the philosophy of mind and the successor to the classical mind-body problem and the more recent body-body problem (Thompson 2007). The body-body problem is the question of how the bodily subject can be at once subjectively lived and an organismic body that is embedded in the world. The body-social problem elaborates on this and is concerned with the question of how bodily and social aspects figure in the individuation of the human individual mind. Philosophers of cognition systematically assume that the mind is essentially embodied, while the social world remains the context in which the embodied mind

is embedded. On this view, the social arguably shapes the mind, but it does not figure in the constitution of the mind itself.

In what follows, I first show that Noë's proposal entails the same presupposition and thus invites a new form of methodological individualism that risks limiting conceptual understanding to the endeavour of an isolated individual subject. I then introduce and discuss an alternative proposal for a model of the individual mind as a *socially enacted self*. I argue that since the world of humans is a world of others and our social relations are what matters most to us, the social must also figure in the constitutive structure of human cognitive individuation.¹ The human mind or self is not only embodied but also genuinely social. From an enactive viewpoint the self can be considered as a self-other generated autonomous system, whose network identity is brought forth through individual's engagement in bodily-mediated social interaction processes of *distinction and participation*. Distinction and participation refer to the two intrinsic goals that the individual follows and needs to balance. Distinction means to be able to exist as individual in one's own right. Participation refers to an openness to others and a readiness to be affected by them. It refers to the sense of self as connected and participating. Both goals are achieved through engaging and relating to others. The processes that constitute the identity of the human mind are therefore not defined in terms of bodily but rather interpersonal relations and interactions. On this enactive approach to the self, the body is not equated with the self but instead seen as that which grounds a double sense of self as a separated identity and as participating. The body mediates the individual's interactions with others (Kyselo 2014).

I outline how the model of the socially enacted self can combine with and elaborate Noë's actionist account of concepts so as to arrive at an even more encompassing view of human un-

1 By saying that sociality matters *constitutively* for the human self, I mean that without continuously relating and engaging in interactions with others, there would be no human self as a whole. The social is not only causally relevant for enacting selfhood, but it is also an essential component of its minimal organisational structure.

derstanding as well as a deeper appreciation of its fragile nature.

2 The risk of crypto-individualism

Noë observes a dichotomy between what he calls the *intellectualist* approach to concepts, the view that concepts are judgments, which is endorsed by Kant and Frege, and the *existential phenomenological* approach, such as that endorsed by Dreyfus, which argues that concepts are usually only used by the novice, and that understanding is otherwise already given through context and situation.² Noë disagrees with both positions. He rejects the idea that concepts are only judgments, fixed and just “out there”, to help us represent the world; yet contrary to the *anti-intellectualists*, Noë also emphasizes that conceptual understanding is not limited to the novice, but “at work wherever we think and perceive and act and talk”. What the existential phenomenologist thereby misses, according to Noë, is that skillful mastery involves learning and development. Noë assumes that, like intellectualism, anti-intellectualism makes the presupposition that concepts are equal to judgments and thus implicitly reduces the mind to a “realm of detached contemplation” (2012, p. 25). For that reason, Noë calls anti-intellectualism *crypto-intellectualist*.

Noë seeks to find an alternative to the two positions by questioning their very fundamentals. Rather than assuming that the world is just given and that everything is already present to us, Noë emphasizes the active contribution of the individual organism (2004, 2009). He proposes that we should adopt a pluralistic approach to concepts, according to which conceptual understanding is basically having the skills required for accessing the world. There are different types or modes of access to the world, including the modes of perception and action, the (inter)personal, and the emotional mode. On this *pluralistic* account, thinking and perceiving

are not very different from one another. Both are “a skillful negotiation with what there is, just another modality of our environment-involving transactions” (Noë [this collection](#), p. 16). From this perspective, judgements belong to a particular mode of access and form part of a broader set of skills of conceptual understanding. Noë then specifies the nature of our access to the world. The world is not just out there ready to be understood. Rather, it always has to be made available and actively brought into view or into “presence”, as Noë puts it. Concepts are the means by which we can achieve this. They are the techniques “by which we secure our contact” with the world ([ibid.](#)). But bringing the world into presence is not a fixed, one-time or uni-directional endeavour. Conceptual understanding involves continuous engagement with the world; it can change and also fail. Noë proposes the notion of *fragility* as a key for understanding conceptual activity as an open and necessarily vulnerable phenomenon, instead of a perfect application of definite representations of the world. In this way, he overcomes the limited view of both the intellectualist and anti-intellectualist perspectives according to which concepts are judgments about an independent world.

One of Noë’s crucial insights is that the traditional dichotomy between an objectively given world and subjectively experienced, internally-processed data about worldly objects can be overcome by grounding all conceptual activity in a broader “common genus”, i.e., skilful engagement with the world. But what is even more important, and in this I think Noë does not actually diverge far from Dreyfus and other existential phenomenologists, is that the established unity of different modes of understanding is not merely a unity in terms of styles of access to the world, but also a unity grounded in the *individual* mind as a whole. But what is that individual mind as whole?

Noë quite clearly presupposes that *we* are not our brains. We understand the world through navigating it with our thinking, skilful sensorimotor body (Noë [this collection](#), 2004). This view breaks with the cognitivist paradigm with regard to the constitutive elements of the

² The existential phenomenological approach refers to phenomenologists such as Heidegger and Merleau-Ponty who investigate the basic structures of human existence. One of their assumptions is that prior to any reflexive understanding, we are already attuned to the world simply through our bodily being in it. Dreyfus calls this pre-reflexive attunement to the world “absorbed coping” (2013, p. 21).

system that does the understanding, and it also breaks with it with regard to the relation of the understanding system to the environment: the system is not passive, but rather active and dynamical. What this elaboration implies, yet does not make explicit, is the fact that conceptual activity is done by a bodily *agent* who understands or has access to the world. After all, conceptual understanding is not just understanding *about* something but always also understanding *for* someone and *by* someone. To argue that thought and perception are unified as modes of access thus presupposes an individual who employs these different modes of access, someone for whom the world can show up. Without an agent that does the understanding, postulating a unification of modes of understanding would not make any sense, as any understanding would remain an action that has neither origin nor actor.

This is a point that [Evan Thompson](#), who is also a proponent of embodied cognition, has already made on some of Noë's earlier work on enactive perception (2007). According to Thompson, while emphasising the role of experiences of *objects*, Noë underestimates the role of subjectivity as such: the "sensorimotor approach needs a notion of selfhood or agency, because to explain perceptual experience it appeals to sensorimotor knowledge. Knowledge implies a knower or agent or self that embodies this knowledge" (Thompson 2007, p. 260). This is where I think Noë's underlying epistemology requires elaboration. Who or what is the individual subject that engages in this fragile endeavour of securing access to the world?

[Thompson](#) provides an insight that can be seen as a major step into the right direction: he proposes addressing the *body-body problem*, i.e., the question of how the agent can be at once subjectively lived and an organismic or sensorimotor body that is embedded in the world (2007, pp. 235–237), by proposing an enactive notion of selfhood. According to this notion, individual agency is defined in terms of *autonomy*. It is seen as a self-organised network of interconnected processes that produce and sustain themselves as a systemic whole—a bounded identity within a particular domain

(Varela 1997; Maturana & Varela 1987). According to [Thompson](#), it is this autonomous self that gives unity to the sensorimotor skills in terms of self-organisation and operational closure (2005, 2007). Operational closure means that some process relations of the autonomous network remain constant despite structural dependence on the environment, i.e., each process within the network is not only enabling but also enabled by some other process. With the production of such a self-organised autonomous identity the individual also acquires a basic subjective perspective, from which interactions with the world are evaluated respectively. This subjective perspective is what [Thompson](#) calls a pre-reflective bodily self-consciousness (2007, p. 261).

On Thompson's enactive account, the individual is now not only active and embodied but also an autonomous subjective agent. Importantly however, Thompson shares with Noë a dubious fundamental pre-supposition, namely the idea that the individual mind or subject can be equated with the individual sensorimotor body or organism. The autonomous agent is a self-organised "sensorimotor selfhood" ([Thompson 2005](#), p. 10). As a consequence, in both Thompson and Noë's views, the mind is empowered and freed, as it is no longer restricted to the passive, information-consuming existence that is distant to the world and confined to the narrow shells of our heads. Nevertheless, it still remains a mind of a body in isolation: in isolation from the world of others.³ This risk of an individualist account of the agent is the first horn of a dilemma underlying Noë's proposal. The second horn has to do with the fact that for Noë understanding is actually *not* an isolated endeavour. The social world is mentioned

3 [Thompson](#) clearly recognises the importance of intersubjectivity for the process of understanding, arguing that "human subjectivity is from the outset intersubjectivity, and no mind is an island" (2007, p. 383). He proposes (in line with Husserl) that humans are from the beginning intersubjectively open. However, it seems that Thompson's emphasis on sociality is either developmentally motivated and concerned with the intersubjectively-open intentionality in object perception or a question of our (rather sophisticated ability) to understand others and to make the distinction between self and other. But the subject herself, despite being intersubjectively open, is still a "bodily subject" ([Thompson 2007](#), p. 382). In other words, the structures of subjectivity itself, the very network processes that bring about the individual as an autonomous system, are determined bodily, not intersubjectively.

throughout the paper in the form of other subjects that seem to enable the individual's understanding in various ways. Some of the skills of access are interpersonal and also, as Noë emphasizes, have to be learned.

The question is, how do we learn skills? We usually learn through a teacher, and thus through the help of another being. Similarly, how do we discover a piece of art? By discussing it with a friend, who helps to bring about a new perspective on it. The person whom we misunderstand and try again to understand is another subject. Understanding is a highly intersubjective endeavour, not only developmentally—in the sense that we need others at some point in life to learn a particular skill—but also in a continuously on-going sense, for much of the very process of human understanding happens through and with others contemporaneously. Strikingly, however, though Noë admits this in acknowledging that understanding happens through communication and thus through the contribution of other subjects, the social does not seem to matter *constitutively* in his general theory of conceptual understanding. The mechanism and structures of the process of understanding are defined in terms of sensorimotor processes, not in terms of interactions with others, and the unity that grounds conceptual understanding is constitutively the sensorimotor body in object-oriented action; it is not, more dynamically put, the individual in its relation to other subjects. The worry is that in Noë's approach, the social part of the world would therefore only play the *weak role* of an outside and divided context. In contrast, on a *strong* reading of the relation between understanding and sociality, engagements and relations with others would have a more than developmental or contextual relevance. Instead, they would also be considered part and parcel of the very structure of the process of understanding, and they would (as I argue below) figure in the minimal constitution of autonomous selfhood.

Noë characterises Dreyfus's anti-intellectualist stance as "crypto-intellectualist" because Dreyfus allegedly accepts the premises of the intellectualist's view that understanding is rule-based judgement. Yet one might say that in his

attempt to overcome the dichotomy between existential phenomenology and classical conceptualism, Noë inherits a very similar problem. Noë's actionist approach opens the individual up to the world; but, perhaps because he is trying to avoid an implication of Dreyfus' existential phenomenology, namely the risk of losing the individual (as already immersed) in the world, Noë also risks over-emphasizing the status of the embodied individual, thereby missing the deeper relation between the individual and the social world. The undesirable implication is that conceptual activity is essentially an isolated undertaking (since according to standard approaches to embodiment there is nothing social about the individual body or organism *per se*). It is the lonesome individual by herself who navigates through the world, equipped with a great set of skills that enable her to act and to secure the access to the world.⁴ Because Noë seems to implicitly accept the individualistic premise of the traditional cognitivist view, one might say that that his proposal is *crypto-individualist*.

Noë is not alone in making the crypto-individualist presupposition. According to Post-Cartesian and non-cognitivist philosophy of cognition, the mind supposedly involves an active and dynamical engagement with the social and material environment, and also has an experiential dimension (Shapiro 2011; Clark & Chalmers 1998; Varela et al. 1993; Thompson 2007). But the integration of these aspects, and in particular that of the social and bodily dimension with regards to the individual that has or is the mind still remains a fundamental question. This is what I have called the *body-social problem*: how can the mind be at once a distinct bodily individual but at the same time remain open and connected to the social world? At the moment there is a dichotomy between views that posit that the mind is embodied and views that emphasize the relevance of situatedness and embeddedness. On the former view, the mind is active but confined to being an isolated indi-

4 Note that it does not actually matter whether one posits that the mind is in the head or in the body, both claims are compatible with the weak reading of the interrelation of individual and social world, according to which the social remains separated from the individual.

vidual. On the latter, the mind is primordially immersed in the (social) world. The first view risks a new form of methodological individualism where the individual mind, while no longer restricted to the brain, is now confined to the body. Here the social world becomes the external, independently given world into which these newly embodied and active, yet essentially isolated individuals parachute (Kyselo 2014).⁵ The second view focuses too much on the interaction dynamics and risks losing the immersed individual mind in the world (and social interactions), thereby blurring the very epistemological target of our philosophical inquiry (Kyselo 2013, 2014).

The body–social problem reveals a deeper linkage between Noë and the stance of the existential phenomenologist that he actually seeks to debunk. Both positions disagree with the traditional Cartesian picture of the mind; both hold that embodiment matters vitally for the mind. But notice that they also focus on different aspects of what a true alternative to the classical view might look like. The overall alternative basically involves a fundamental shift in thinking about the relation between an individual and the world. In this vein, Noë is right to emphasise the individual's power, giving it more responsibility in the very construction of its own mind and of the world it experiences, but so are the existential phenomenologists when they focus on worldly embeddedness and the fact that a great deal of our being in the world relies on pre-given structures that can surpass the individual's capacities. An emphasis on individual action and responsibility cannot mean that the individual is all alone. We would not have made enough progress if the main difference between Noë's proposal and the representationalist division between individual and world was that now, while being able to move towards the world, the world does not also move toward us but remains separate with regard to other subjects. Other people are active, too, and they shape not merely the world for us but also

who we are as subjects. But, speaking to the potential worry of losing the individual in worldly engagements, the solution is of course neither to negate any need for differentiation nor the necessity of the individual to have its own share in the very mechanism of understanding the world. Where I think both positions go wrong is in extrapolating from a part of adult human phenomenology (even when it is paired, as in Noë's case, with an objective account of the constitutive mechanism of experience) to a general theory of understanding. In crypto-individualism the individual mind carries a heavy burden. It is free from passivity and yet enormously restrained by the responsibility of achieving the access to the world (and the social world) and itself, all by itself. Existential phenomenologists, in emphasising the importance of the social world and its pre-given structures in bringing about understanding then ease the burden and free the individual from some of the responsibility in achieving this; and yet at the same time they also risk depriving the individual of its power and right to have a say in that endeavour.⁶

It should be clear that neither position on its own will suffice to overcome the dichotomy inherent in the intellectualist view on concepts. The individual cannot understand the world simply by being an individual body, but neither is the world already understood just by simply being immersed in it.

3 Deep dynamics and the enactive self

There exists a middle ground from which the dilemma of having to choose between too much or too little individualism can be avoided and a more complete epistemological basis for conceptual understanding achieved. Finding this middle ground basically consists in re-thinking the nature of the mind and of human understanding while doing more justice to the deep interrelation between individual and social world. To this end I have recently proposed the

⁵ This image is adapted from Varela et al. (1993), who criticise the traditional view as implying that the environment is a "landing pad for organisms that somehow drop or parachute into the world" (p. 198); instead, they argue that the relation between world and individual mind is co-determining.

⁶ This commentary is not the place to discuss this issue in detail, but it should be noted that such a view can be expanded to political philosophy and the philosophy of law, where it might have far reaching consequences for questions concerning the nature of individual rights and approaches to legal responsibility.

concept of the *socially enacted self* (Kyselo 2014, 2013; Kyselo & Tschacher 2014). On this approach, the individual is not sufficiently determined in terms of active embodiment; instead it is thought to incorporate social and relational processes into the structure that makes up its identity as an individual. This suggests that without a “social loop” we cannot speak about the human self as a centre of individuation in any interesting sense. After all, humans do not merely distinguish themselves against a background of material objects, but, crucially, against the world of other humans. They become someone, an identifiable individual against a world of other individuals and social groups.

This idea should become clearer by reconsidering, or making more explicit, a number of insights already implied in diverse approaches in embodied cognitive science.

First, Noë’s crypto-individualism captures something essential about the ways humans access the world: we often experience the process of understanding as something we do by ourselves—the concepts we acquire and employ are ours and to a large extent we appear to be in control in our attempts to secure the world. Noë’s other important insight is that conceptual understanding is an achievement. It is a far-from-perfect endeavour, involving experiences of vulnerability, openness, of not always being able to own and to access the world.

The second insight is appreciated in the debate on extended cognition. Clark & Chalmers in their now classical paper “The Extended Mind” propose that a tool, such as a notebook or a computer, can count as part of the individual mind (1998). This essentially functionalist position goes against Noë and “beyond the sensorimotor frontier” (Clark 2008, p. 195)—the mind is not restricted to the body but spreads across neuronal, bodily, and environmental features. The extended cognition approach to embodiment has been criticised for being too liberal, since it lacks both a principled definition of “body” and of “cognition”. It remains unclear how an environmental prop or technology could be integrated into the cognitive architecture of an individual mind (Kyselo & Di Paolo 2013, see also Menary this collec-

tion). Yet, despite these shortcomings I believe there are two important insights in this extended functionalist account: first, that the individual should not be restricted to the biological realm (be it the brain *or* the body) but incorporates tools and technologies, and second, that the mind transcends the individual physiological body and that the world matters constitutively for determining the boundaries of the mind.

The third insight comes from the enactive approach to cognition, which proposes that the mind is basically an autonomous system that self-organizes its identity based on operational closure. The enactive approach thereby shares with extended cognition the idea that the individual is not clearly separable from the environment. On the enactive view, the individual’s mind is “defined by its endogenous, self-organizing and self-controlling dynamics, does not have inputs and outputs in the usual sense, and determines the cognitive domain in which it operates” (Thompson 2007, p. 43). Identity is therefore not a given thing or a property, but *relational*: brought forth through the individual’s on-going and dynamical interaction with the world. This approach adds an insight derived from philosophy of biology, namely that like living beings, cognitive beings create an identity that they strive to maintain, and that understanding the world depends on the purposes and concerns of that identity (Weber & Varela 2002; Thompson 2007) in that they guide and structure our understanding.⁷

The three variants of embodied cognitive science therefore all reject the mind–body dichotomy and emphasise a dynamical interrelation between embodied individual and world. All of them however, either miss or do not fully acknowledge that the world is social and that the individual is also a psychological and social being whose concerns are more than object-oriented. This is where the enactive approach to the social self comes into play. It basically elaborates on and integrates the above insights, i.e., action (sensorimotor cognition), co-constitution

⁷ Interestingly, this is also an insight Dreyfus pointed out much earlier when he argued that the “human world, then is prestructured in terms of human purposes and concerns in such a way that what counts as an object or is significant about an object already is a function of, or embodies, that concern” (1972, p. 173).

(extended cognition), and grounding in selfhood (enactive cognition), by adopting a much more radical perspective on the dynamical interrelation between the individual and the world—let us call this perspective *deep dynamics*. Deep dynamics means that the nature of the relation between individual and world is one of strong co-constitution: not only does the individual actively shape and structure the world, the world, too, affects the individual in its basic organisational structure. If identity and domain depend on each other in a strong and mutual sense, as the enactive approach to cognition has it, then even more advanced non-organismic or virtual notions of the body do not change the fact that the organismic bodily domain is an individualist domain (Kyselo & Di Paolo 2013). In other words, the organismic body cannot be related to the social at the same level of organisational closure. The enactive approach to the self would suggest instead that the level at which human selves can be usefully operationalised as autonomous identities is social, not merely embodied. Admittedly, by emphasising how conceptual understanding is shaped through social engagements with others, Noë's approach obviously also implies a bi-directional relation between individual and world. Similarly, as we have seen above, Thompson's sensorimotor subject is also clearly involved in intersubjective interactions (2005, p. 408). However, the bi-directional impact in these accounts is more *shallow* than in the present proposal, as they consider the (social) world to play a contextual or developmental role, or to matter with regards to shaping object-recognition. In deep dynamics, in contrast, we expand on the insight of extended cognition that the mind transcends brain and body by acknowledging that this not only the case through interactions with tools but also through our social interactions and relations with other subjects. The idea then is that *qua* being embedded in a social world, the self, and by that I mean the individual as a whole, constitutively relies on its interactions and relations to other subjects. According to this elaboration on the enactive account of selfhood, the self can be defined as a socially enacted autonomous system. It is:

a self-other generated network of precariously organized interpersonal processes whose systemic identity emerges as a result of a continuous engagement in social interactions and relations that can be qualified as moving in two opposed directions, toward emancipation from others (distinction) and toward openness to them (participation). (Kyselo 2014)

In line with the concept of operational closure, both types of processes, distinction and participation, are required to bring about the individual self. Without distinction, the individual would risk immersion or becoming heteronomously determined and forced to rely on the next best or a limited set of social interactions. But without participation and an act of openness towards others, the individual eschews structural renewal, thus risking isolation and rigidity (Kyselo 2014). The point, however, is that this form of operational closure contains social interactions. In enactive terms, this is to say that the individual is at the same time self-and-other-organized. As a consequence, the self is not a given nor an individual bodily achievement but also and necessarily co-constructed with others. Both the individual and the world (that is, other subjects) have a say in the constitutive mechanism of someone's mind. In contrast to Noë's presupposition, the mind cannot be equated with the active body. Rather, the sensorimotor body becomes the ever-evolving interface that in being with others co-generates the very boundaries of what we call the self (Kyselo 2014).

At this point, proponents of embodiment might still want to insist that there is something about the body's role in grounding the sense of self that non-negotiably remains entirely independent from social interactions. I agree, if by "sense of self" one refers to the self as mere biological identity. However, if by "self" we mean the human self in distinction from other humans, then the proposed view challenges this intuition. It does this, however, without giving up the insight that the self has to do with individuation. The enactive notion of autonomy and self-organization saves the indi-

vidual from immersion in the social world by appreciating that the distinction between individual and world is an organisational, not ontological distinction. Our sense of being a distinct someone is something that is achieved together with others, not just qua being a biological body.

The basic idea of the socially enacted self is therefore not to overcome the tension entailed in the body-social dichotomy but rather to welcome and recognise it as a necessary property of mind itself and to thus integrate this tension into a general theory of understanding. On this view, the individual mind has to continuously negotiate its identity as an individual agent and its understanding in dependence on other subjects. As a consequence, uncertainty, conflict, and a permanent need for negotiation and co-negotiation are part and parcel of being an essentially social human mind. This is why it might be useful to distinguish several senses of fragility. Fragile understanding is one of them. But on the enactive account of selfhood, mind itself is fragile.

4 Varieties of co-presence

Let us now explore a couple of implications that a deep dynamics view has for conceptual understanding. By basing conceptual understanding on an understanding of the individual as a socially enacted autonomous system, we can do justice to existential phenomenologists who emphasize the importance of situatedness and flow and also to Noë's rightful actionist call for emancipation of the passive individual mind. For Noë, the unity of conceptual modes is derived from positing an active, thinking, sensorimotor body. The present proposal suggests that the unity is grounded in a socially co-organized individual. Noë's idea of thinking of experiencing and understanding the world as a "relation between a skillful person and really existing thing" (2012, p. 42), could thus be elaborated by saying that the intentional relation is also a relation to other subjects, so that intentionality is actually co-generated. Yet this co-generated intentionality is not merely about sharing a perspective on the world; it is a co-generated rela-

tion that feeds into the very organisational structure of mind itself. The person involved in the intentional relation is a social subject. In accordance with the two-fold structure of socially enacted autonomy, this would also mean that self-reflexivity has a social structure, entailing a sense of being a self as separate individual and a sense of being open and connected to the world.

Here lies the deeper reason for why the process of understanding is fragile. The fragility of understanding consists precisely in the fact that the unity of mind is never a given, but is itself an on-going achievement. Since, as I suggest, this is an achievement with others, presence does not merely depend on what we do, but also on what others do, and especially on what we do with them. In other words, presence is actually *co-presence*. It is clearly outside the scope of this commentary to explicate this in more detail, but generally speaking it means that understanding simply never really is the endeavour of an individual mind. This complements Noë's perspective and invites future explorations in at least two fundamental senses.

First, with regards to the role of others in empowering the individual by enabling access to the world: our conceptual skills are acquired and the acquisition of these skills usually happens in interaction and by learning together with others. But our ways of understanding are also *continuously* shaped and mediated by being with others, be it through cultural norms, biases, advice, or advertisement. Apart from the obvious fact that much of instantaneous understanding happens together with others, even in the absence of others, in the process of understanding, we often presuppose another subject or at least some implicit act of relationality. Noë says that "there is no such thing as a perceptual encounter with the object that is not also an encounter with it from one or another point of view" (2012, p. 138). I could not agree more, and yet I suggest we also embrace the idea that these other viewpoints are not merely defined in terms of changes in head or body-movement but also in terms of loops to and from different subjective and intersubjective view points.

If conceptual understanding has the purpose of bringing us into contact with the world, as Noë claims, then we should not underestimate the role of others and of our being open to them in making this contact possible. To consider human understanding as fragile is also to admit a limitation of the individual's capacities and to allow others and our dialogues with them to play a fundamental role. In this sense fragility can be a source of power. Our minds are open, not only to the world, but also to contributions from others.

But that said, and this is the second and final implication of the enactive self for the basic nature of human understanding, the social nature and fragility of mind also restricts the individual's capacities. When the social plays a marginal and contextual role, the individual's responsibility in understanding the world is immense and the optimism in the individual's capacities can become a heavy burden. The other side of fragility is that the presence of the world is not only "not for free", as Noë puts it, but it is actually sometimes not available at all. It is not available because other subjects have a say in the construction of our understanding, and given that they have perspectives and interests of their own, their contribution may sometimes be out of reach, run contrary to what we need, or even confuse us deeply. The fragile nature of our social mind can therefore also deny us access to the world.

5 Conclusion

In his book *Varieties of Presence*, Noë refers to Kafka's *The Metamorphosis* (1915), the story of Gregor Samsa, who wakes up as an insect, lying on his back, unable to move. Noë uses the story to illustrate the upshot of his philosophy of understanding. "We are not only animals", he says, but we "achieve the world by enacting ourselves. Insofar as we achieve access to the world, we also achieve *ourselves*" (Noë 2012, p. 28).

On the presented alternative, the actionist nature of self-achieved understanding is only half of the story. I have suggested that our minds and selves are genuinely social and thus transcend the limits of our bodily existence. The human self vitally depends on others and is

achieved together with them, through negotiating a permanent tension of maintaining a sense of individuality while not losing the connection to others (distinction and participation).

From this perspective, the point of Kafka's story is therefore not so much to deny that we are animals, but rather to claim that we are *social* animals that achieve ourselves *together* with others. Reflecting the basic insight of this paper, the story thus illustrates the fragility and social nature of human existence. It is an expression of desperation and of the suffering that can come when others refuse or are unable to comply with our basic needs: being recognised as individual and as someone who belongs to others. Having lost contact with himself as a human subject in the bureaucratic machinery of his professional life, Samsa awakes as an insect, his new embodiment an imprint of alienation and loss of recognition. But the loss cuts even deeper. With his alien embodiment Samsa the insect is rejected by his family, so that he finds no salvation in his private life. Samsa dies from social isolation. From an enactive view of the self as a joint achievement, Kafka's *The Metamorphosis* captures (like much of his other work) the consequences of our deep vulnerability and limited freedom and the drama of the loss from which we can suffer precisely because we are social beings.

The social structures that we depend upon empower our ways of understanding; yet for the same reason they can also enslave us, and seriously limit our mental capacities. This, I suggest, is not merely the case for institutions and their bureaucratic apparatus but also applies to our direct intersubjective relations, be they with lovers, friends, family, or co-workers.

Presence is therefore not simply availability—since this would suggest the subject's unwarranted access to the world. Presence is rather a joint achievement, and the nature of doing things together is that there will always be leaps and limitations. In this way, failure and limited control over the ways we understand the world are not entirely the responsibility of the individual and its techniques and skills, but also a deeper expression of the genuinely social and co-constructed nature of understanding.

Acknowledgements

I would like to thank Gabriel Levy and Mike Beaton as well as two anonymous reviewers for their useful comments. My gratitude also goes to the editors and organisers of the MIND group, Jennifer Windt and Thomas Metzinger. The MIND-group has been a unique source of inspiration and support. This work is supported by the Marie-Curie Initial Training Network, “TESIS: Toward an Embodied Science of Inter-Subjectivity” (FP7-PEOPLE-2010-ITN, 264828) and by the “Science Beyond Scientism” Research Project at VU University of Amsterdam.

References

- Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. Oxford, UK: Oxford University Press.
- Clark, A. & Chalmers, D. (1998). The extended mind. *Analysis*, 58 (1), 7-19. [10.1111/1467-8284.00096](https://doi.org/10.1111/1467-8284.00096)
- Dreyfus, H. L. (1972). *What computers can't do*. New York, NY: Harper and Row.
- (2013). The myth of the pervasiveness of the mental. In J. K. Schear (Ed.) *Mind, reason, and being-in-the-world: The McDowell-Dreyfus debate* (pp. 15-41). London, UK: Routledge.
- Kafka, F. (1915). The Metamorphosis.
- Kyselo, M. (2013). Enaktivismus. In A. Stephan & S. Walter (Eds.) *Handbuch Kognitionswissenschaft* (pp. 197-202). Stuttgart, GER: J.B. Metzler.
- (2014). The body social: An enactive approach to the self. *Frontiers in Psychology*, 5. [10.3389/fpsyg.2014.00986](https://doi.org/10.3389/fpsyg.2014.00986)
- Kyselo, M. & Di Paolo, E. (2013). Locked-in syndrome: A challenge for embodied cognitive science. *Phenomenology and the Cognitive Sciences*, 3 (1), 1-26. [10.1007/s11097-013-9344-9](https://doi.org/10.1007/s11097-013-9344-9)
- Kyselo, M. & Tschacher, W. (2014). An enactive and dynamical systems theory account of dyadic relationships. *Frontiers in Psychology*, 5 (452). [10.3389/fpsyg.2014.00452](https://doi.org/10.3389/fpsyg.2014.00452)
- Maturana, H. R. & Varela, F. J. (1987). *The tree of knowledge: The biological roots of human understanding*. Boston, MA: Shambhala Publications.
- Menary, R. (2015). Mathematical cognition. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Noë, A. (2004). *Action in perception*. Cambridge, MA: MIT press.
- (2009). *Out of our heads: Why you are not your brain, and other lessons from the biology of consciousness*. New York, NY: Hill and Wang.
- (2012). *Varieties of presence*. Cambridge, MA: Harvard University Press.
- (2015). Concept pluralism, direct perception, and the fragility of presence. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Shapiro, L. (2011). *Embodied cognition*. New York, NY: Routledge.
- Thompson, E. (2005). Sensorimotor subjectivity and the enactive approach to experience. *Phenomenology and the Cognitive Sciences*, 4 (4), 407-427. [10.1007/s11097-005-9003-x](https://doi.org/10.1007/s11097-005-9003-x)
- (2007). *Mind in life: Biology, phenomenology, and the sciences of mind*. Cambridge, MA: The Harvard University Press.
- Varela, F. J. (1997). Patterns of life: Intertwining identity and cognition. *Brain and Cognition*, 34, 72-87. [10.1006/brcg.1997.0907](https://doi.org/10.1006/brcg.1997.0907)
- Varela, F. J., Thompson, E. & Rosch, E. (1993). *The embodied mind: Cognitive science and human experience*. Cambridge, MA: MIT Press.
- Weber, A. & Varela, F. (2002). Life after Kant: Natural purposes and the autopoietic foundations of biological individuality. *Phenomenology and the Cognitive Sciences*, 1 (2), 97-125. [10.1023/A:1020368120174](https://doi.org/10.1023/A:1020368120174)

Beyond Agency

A Reply to Miriam Kyselo

Alva Noë

In this paper I respond to [Kyselo's \(this collection\)](#) claim that actionism, and other versions of the enactive embodied approach to mind, fail to accord social relations a constitutive role in making up the human mind. I argue that actionism can meet this challenge—the view makes relations to others central to an account of human experience—but I also question whether the challenge is clear enough. I ask: what exactly does it mean to say that social relations play this sort of constitutive role?

Keywords

Actionism | Body-social problem | Concept pluralism | Concepts | Consciousness | Enactive account | Enactive self | Evans | Fragility | Frege | Individualism | Intellectualism | Kant | Organized activity | Perception | Plato | Presence | Sensorimotor account | Socially enacted autonomy | Socially extended mind | The intellectualist insight | The intellectualist thesis | Understanding | Wittgenstein

Author

[Alva Noë](#)

noe@berkeley.edu

University of California,
Berkeley, CA, U.S.A.

Commentator

[Miriam Kyselo](#)

miriam.kyselo@gmail.com

Vrije Universiteit
Amsterdam, Netherlands

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

In my contribution to this volume ([Noë this collection](#)), I seek to bring out the truth in intellectualism. The intellectualist is right, I concede, that understanding is at work throughout the domain of agency—wherever we can talk of perception, or thinking, or action. Understanding is pervasive. The trouble with intellectualism, I argue, is that it cleaves to an unrealistic conception of what is demanded for understanding to come into play. In particular, it adheres to an over-intellectualized conception of understanding, according to which an action, or a perception, can be conceptual only if it is guided, as it were from above, by explicit acts

of judgment. In my target paper I also criticize *anti*-intellectualist views, such as that of Dreyfus, for failing to break with intellectualism; such views reject the pervasiveness of the understanding because they accept the intellectualist's hyper-intellectualized conception of what understanding is and because they find it implausible that our experiential or cognitive lives are intellectual in this way. In this brief reply to Kyselo's excellent commentary, I would like to say something about what the anti-intellectualism of the sort I criticize in the paper *gets right*. I now want to try to bring out the insight in anti-intellectualism.

2 The truth in anti-intellectualism

If the intellectualist is right that understanding saturates the space of agency, the anti-intellectualist is right that there is also understanding *beyond* the limits of our agency. Stanley (2011, cited in Noë [this collection](#)) relied on the opposition between the personal and the subpersonal; he supposed that what makes a mere reflex, which is subpersonal, an action, which is personal, is that it is guided by knowledge or reason. But the opposition between reflex and action is not exhaustive, and the crucial dimension is not that of the contrast between the personal and the subpersonal. Consider conversation, as an example. We can characterize conversation as a personal-level action. But there is a way of describing the phenomenon that defies such characterization. When two people talk they adopt similar postures, they pause at coordinated intervals, they adjust their volumes to match each other, they move their eyes and modify their dialects, all in ways that are governed by their interaction (see Shockley et al. 2009 for a review of this literature). Talking is what I elsewhere call an “organized activity” (Noë [in press](#)). One remarkable feature of organized activities, in this sense, is that they are not guided by the participants or authored by them. Another is that they are carried on spontaneously and without deliberate control. And yet another is that they are clearly domains in which highly sophisticated cognitive capacities—looking, listening, paying attention, moving, undergoing—are put to work.

Notice: I said above that talking, in the sense I have in mind, is not a personal-level activity. What I mean by this is that the sort of tight coupling and temporal dynamics, the sort of organization we see at work when people talk, is not best characterized at the level of minutes, hours, choices, etc. that normally characterize the personal level. But nor is this a phenomenon of the subpersonal level. For one thing, we aren’t interested in something happening in the nervous system of *one* individual. We are interested in something encompassing two (or more) people. For another, we aren’t interested in processes unfolding at time-scales of

milliseconds. No. We *are* interested in what people do, but in a manner that is truly *beyond* agency. We are interested, here, in a phenomenon of the *embodiment level* (as distinct from the subpersonal or the personal level).

And yet we remain, when thinking about conversation—or any other organized activity—very much in a domain where we can and must speak of cognitive achievement, understanding, skill, and so on.

One upshot of these considerations, then, is that while understanding, as I argued above, is a necessary condition of agency, it is also present beyond its limits. Another is that understanding beyond the limits of agency cannot be understood individualistically. This is obvious in the case of intrinsically social activities, like conversation, but it is also true for organized activities that can be carried out by solitary individuals (such as *seeing*, for example).

The thing that anti-intellectualism gets right, as I see it, is the appreciation that a great deal of what we *do*, isn’t really done *by us*: activity happens to us; we find ourselves organized. We are made what we are in the setting of organized activities.

From the standpoint of the theory of organized activities—presented in more detail in Noë ([in press](#))—we are creatures who are from the very beginning caught up in world and other-involving organized activities; these activities form the lived substrate of our biographical lives as persons. Actionism, in these ways, is committed to a radical form of anti-individualism.

3 The challenge of crypto-individualism

Now, Kyselo has criticized actionism not for ignoring the social, but for failing to treat the social as constitutive of human cognitive organization. Kyselo’s point is that for actionism, other people and our relations to them “shape” the mind, but they do so in the same the way that any environmental conditions cause, constrain, or enable human experience; the view makes no allowance for the stronger possibility that other people and our social relations with them are actually *constitutive* of what it is to be a human

being. So she writes, with actionism as one of her targets in mind:

Philosophers of cognition systematically assume that the mind is essentially embodied, while the social world remains the context in which the embodied mind is embedded. On this view, the social arguably shapes the mind, but it does not figure in the constitution of the mind itself. (Kyselo [this collection](#), p. 2)

And she goes on to explain:

I argue that since the world of humans is a social world of others and our social relations is what matters most to us, the social must also figure in the constitutive structure of human cognitive individuation. The human mind or self is not only embodied but also genuinely social. (*ibid.*, p. 2)

In a footnote, she then elaborates:

By saying that sociality matters constitutively for the human self, I mean that without continuously relating and engaging in interaction with others, there would be no human self as a whole. The social is not only causally relevant for enacting self-hood, but it is also an essential component of its minimal organizational structure. (*ibid.*, p. 2)

Now, I admit that the language of earlier work (Noë 2004, 2012) can be taken to suggest something like crypto-individualism. In so far as I talk about presence as something that thinkers and perceivers “achieve,” and in so far as I insist that, in achieving the world’s presence in thought and experience, we also achieve ourselves, it can perhaps sound like I am describing the enactive feats of a heroic solitary agency.

I admit that’s how it sounds. But I was careful to warn against being misled in this way. So, for example, in a passage immediately following one that Kyselo cites, I write:

But we are not only animals. I am also a father, and a teacher, and a philosopher, and a writer. These modalities of my being were no more given to me than my ability to read and write. I achieve myself. Not on my own, to be sure! And not in a heroic way. Maybe it would better to say that my parents and my friends and family and children and colleagues have achieved me for me. The point is that we are cultivated ourselves—learning to talk and read and dance and dress and play guitar and do mathematics and physics and philosophy—and in this cultivation worlds open up that would otherwise be closed off. In this way we achieve for ourselves new ways of being present.

Here I explicitly repudiate heroic individualism; we achieve ourselves with and through others; we are cultivated by a world full of others and that’s the setting in which we bring the world into focus for consciousness.

Perhaps another feature that feeds the appearance of crypto-individualism is the availability of an idealist or anti-realist reading of enacting or achieving presence. It is not in fact my view—Kyselo herself is clear about this—that we make the world, or construct it. The world shows up for us, in perception, and in thought, and for action. But it doesn’t show up for free. Just as you can’t encounter what a text means if you don’t know how to read, so you can’t see what is there to be seen without the battery of understandings necessary for reaching out and picking it up.

We don’t make the world, just as we don’t make other people. In fact, the world, and others, are necessary for us to achieve contact with it *in three distinct ways*. First, our experience of others and the world depends on their existence. If they weren’t there, we couldn’t achieve access to them. Second, our possession and exercise of the relevant skills may require the presence and participation of others. Think of the turn-taking dance that is conversation; you can’t do that without the other. Third, our possession of perceptual and cognitive skills of access de-

depends on our development in the setting of personal relationships.

Does the commitment of actionism to these three kinds of dependence of our experience on our engagement with others meet the standard of offering an account of other people as not merely shaping but as constituting our mental lives? If not, I hope to be told why.

Let me offer a final example to try to clarify what is at stake. Take a baseball team. There will be nine players on the field at a given time during a game: a pitcher and catcher, three basemen, a shortstop, and the three outfielders. Notice that there are two different ways in which we can individuate these players. We can pick them out by the role that they play—by their position, in baseball parlance—or we can pick them out by *the player*, that is, by the particular person who is playing the role. Take the shortstop, for example. The shortstop is the near outfielder, or the far infielder; he is positioned between 2nd and 3rd bases. His job is to field balls hit to him and to deliver the balls to teammates in ways that work to his team's advantage. For our purposes it is important to notice that a shortstop is a social creature in the sense that a) to be a shortstop is to play a role that can only be specified by naming other positions and shared goals and needs, and b) that there is no such thing as a shortstop outside of the context of convention, practice, and history—for that is what baseball is: a structure in a temporally extended space of convention and practice. A shortstop, we might say, is a thoroughly social kind of thing. It is constituted by social relations.

Notice that this way of thinking about what it is to be a shortstop takes nothing away from the fact that shortstops are embodied and that they are in continuous dynamic exchange with their physical environment. The quality of a shortstop is usually framed in terms of the range of ground he can cover, the softness of his hands, the strength of his arm, the delicacy and control of his footwork, and finally, his understanding of what to do in the split-second heat of play. Physical and intellectual skill are all properties of this essentially social being, the

shortstop. And this is so for all the other players.

Now, the fact that being a shortstop is something “whose identity is brought forth through body-mediated social interaction”, as we could say, borrowing Kyselo's words ([this collection](#), p. 2), doesn't entail that the flesh-and-blood human being who is playing shortstop is also in the same way identity-dependent on his or her social relations. The individual existence of the man, after all, the actual guy, the living human organism, is presupposed by his entering into the kinds of relationships that can make it the case that he is also a shortstop.

This sort of consideration can be generalized: just as we can distinguish the player from the position he plays, so we can distinguish the *human being* from the *person* he or she also is. Personhood is enacted, achieved, or performed in ways not so different from the way being a baseball-player is undertaken. A person is defined by nesting and overlapping roles—daughter, employer, citizen, rebel, lover, failure, and so on. And these roles are genuinely constitutive of who or what a person is, of his or her identity. Truly these constitutive features that make a person the person she is are robustly and thoroughly social, in all the ways being a shortstop is social. You can't be a person on your own, any more than you can be a shortstop on your own. Persons are creatures of normative, evaluative spaces. Persons are performers. They perform their personhood. And they bear the ever-present burden of being evaluated. That, finally, is the difference between mere action and performance. Performance, as distinct from mere action, happens against the background of the possibility of being judged (good dancer, good father, good lover, good student, etc.).

Personhood is enacted. But what about being human? Is that enacted as well? Is one's status as a human being, like one's status as a person, or a shortstop, something that is accomplished through one's body-mediated social interactions?

This much is clear. Being a distinct human being is antecedent to entering into the kinds of

relationships that constitute one's being a person, or a shortstop. So it can't be that it is the same kinds of relations with others that constitute one's personal identity (in my sense) that constitute one's organismic identity as a human being. My question for Kyselo, then, would be: why should we say that human beings, above and beyond the persons they enact, are, in the relevant sense, constitutively social? Or better still, the question is: what is the relevant sense of "constitutively social"?

Let me be clear that I think it would be a mistake to hold that personhood, bound up with practice, convention, and history, though it is, is *merely* cultural, and that this cultural structure is stamped or imposed onto a pre-given biological substrate (the human being). No, each of us is both a human being and a person and any comprehension of our nature needs to do justice to both of these. A biological theory of *us* will be a theory of creatures who are both persons as well as organisms and will take seriously the way these loop back and down and the way they interact.

4 Conclusion

There is much in Kyselo's excellent response to which I have said nothing in reply. I am struck, in particular, by her powerful handling of the concept of fragility. I have tried, in this reply, to show that actionism, despite appearances of heroic individualism to the contrary, recognizes that people spend their lives in worlds that are always ineliminably social.

References

- Kyselo, M. (2015). The fragile nature of the social mind. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Noë, A. (2004). *Action in perception*. Cambridge, MA: MIT Press.
- (2012). *Varieties of presence*. Cambridge, MA: Harvard University Press.
- (2015). Concept pluralism, direct perception, and the fragility of presence. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Noë, A. (in press). *Strange tools: Art and human nature*. New York, NY: Farrar Straus and Giroux.
- Shockley, K., Richardson, D. C. & Dale, R. (2009). Conversation and coordinative structures. *Topics in Cognitive Science*, 1 (2), 305-319.
[10.1111/j.1756-8765.2009.01021.x](https://doi.org/10.1111/j.1756-8765.2009.01021.x)
- Stanley, J. (2011). *Knowing how*. New York, NY: Oxford University Press.

How Does Mind Matter?

Solving the Content Causation Problem

Gerard O'Brien

The primary purpose of this paper is to develop a solution to one version of the problem of mental causation. The version under examination is the content causation problem: that of explaining how the specifically representational properties of mental phenomena can be causally efficacious of behaviour. I contend that the apparent insolubility of the content causation problem is a legacy of the dyadic conception of representation, which has conditioned philosophical intuitions, but provides little guidance about the relational character of mental content. I argue that a triadic conception of representation yields a more illuminating account of mental content and, in so doing, reveals a candidate solution to the content causation problem. This solution requires the rehabilitation of an approach to mental content determination that is unpopular in contemporary philosophy. But this approach, I conclude, seems mandatory if we are to explain why mental content matters.

Keywords

Content determination | Mental causation | Mental representation | Resemblance

Author

Gerard O'Brien

gerard.obrien@adelaide.edu.au

University of Adelaide
Adelaide, Australia

Commentator

Anne-Kathrin Koch

Anne-Kathrin.Koch@gmx.de

Johannes Gutenberg-Universität
Mainz, Germany

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction: The content causation problem

Philosophy delights in those aspects of the world that initially seem obvious and natural, but which on reflection turn out to be deeply mysterious. The mental causation of behaviour is one such phenomenon. Nothing could be more obvious than that our minds matter—that our beliefs and desires, and our perceptions and thoughts ultimately have a causal impact on our behaviour. And yet it has proved notoriously difficult to explain just how this could be the case.

The problem of mental causation has morphed and fragmented over the years. In its

original guise, it was the problem of how a non-physical mental substance or property could causally interact with the physical brain. The obvious solution to this version of the problem was to adopt a thorough-going materialism of some kind, with the consequence that mental phenomena are identified with properties of the brain from which they inherit their causal efficacy.

With the advent of functionalism in the later years of the last century, this “obvious” solution ran into difficulties. If mental phenomena are multiply-realizable, as the orthodox

construal of this metaphysical position seems to imply, then mental properties can't be identified with properties of the brain after all; and since the latter do all the causal work insofar as behaviour is concerned, the problem of mental causation re-asserts itself in a different form (Kim 1992; Crane 1995). This version of the problem of mental causation, which seems to generalise beyond the realm of the mental to all multiply-realizable phenomena, is still keenly debated in philosophy (Kim 2000, 2005; Hohwy 2008).

There is yet another rendering of the problem, however, that revolves around the causal efficacy of the specifically *representational* properties of mental phenomena. This third version typically arises in the philosophy of mind from the conjunction of three widely accepted theses about mental phenomena and their physical realization in the brain:

The content causation problem

1. Mental phenomena are causally efficacious of behaviour in virtue of their representational contents.
2. The representational contents of mental phenomena are not determined by the intrinsic properties of the brain.
3. The brain is causally efficacious of behaviour in virtue of its intrinsic properties.

The first of these theses is a fundamental tenet of both folk psychology and the computational theory of mind that has been constructed on its foundations. It is simply common sense that our perceptions and thoughts are about various aspects of the world in which we are embedded. It is also commonsense that mental phenomena causally interact with other mental phenomena and bodily behaviour in a fashion determined by their *content*—i.e., how they represent the world as being. Fodor refers to this as the “parallelism between content and causal relations” (1987).

The second thesis is widely accepted because most contemporary philosophers think that the representational properties of mental phenomena are determined at least in part by

factors beyond the brain. This is the conclusion drawn from a number of famous thought experiments implicating twin-earth, arthritis, and various species of tree (Putnam 1975; Burge 1979, 1986). But, even more importantly, the second thesis seems to be an entailment of the most popular approach among philosophers for explaining how the representational properties of mental phenomena are determined. This is the conjecture that mental phenomena are contentful in virtue of their causal relations with those aspects of the world they are about (Adams & Aizawa 2010).

The final thesis is consistent with all we know about the brain basis of behavioural causation. While the brain enters into complex causal relations with aspects of the environment via multifarious sensory channels, our best neuroscience informs us that the changes to musculature that constitute our behavioural responses are wholly determined by the intrinsic properties of the brain to which they are causally connected.

In conjunction, these three widely accepted theses form an inconsistent triad. This generates a distinct and narrower version of the problem of mental causation: How can mental phenomena be causally efficacious of behaviour in virtue of their representational contents if these contents are not determined by intrinsic properties of the brain? In what follows, I shall refer to this as the *content causation problem*. This is the version of the problem of mental causation with which I shall be concerned in this paper (see e.g., Kim 2006, pp. 200–202).

There are some philosophers who seek to resolve the content causation problem by rejecting either the first¹ or the third² of the theses composing the inconsistent triad. However, the most popular response has been to reject or at

1 This, for example, is one way of construing Dennett's instrumentalism (1978, 1987).

2 There has been a some discussion in the literature about whether the *relational* properties of brain states are implicated in the causation of behaviour. The standard way of defending this claim is by individuating behaviour *broadly*, so as to incorporate factors beyond bodily movements (Burge 1986; Wilson 1994). But many philosophers think this form of individuation does great violence to scientific practice in general and to neuroscience in particular, and hence this way of resolving the problem of content causation is thought to seem very unpromising (Fodor 1987).

least modify the second thesis. This leads to the the *narrow content program*:

The project of developing an account of mental phenomena according to which (at least the causally relevant component of) their representational properties are determined by intrinsic properties of the brain.

There are a number of different proposals about narrow content in the literature. Two of these have been particularly prominent. One is Fodor's suggestion that narrow contents can be unpacked as "functions from contexts to truth conditions" (1987, Ch. 2). The other is that narrow content is determined by "short-armed functional roles" (Block 1986; Loar 1981, 1982). But these (and other) proposals have been roundly criticised for failing to capture the *relational* character of mental content:

The main charge has been that narrow content, as construed in these accounts, is not real content. When one thinks of an apple, what one thinks about is not a role or a function, but a fruit. Real content must put the subject in cognitive contact with the external world. [...] A water concept, for example, must involve a relation between the thought wherein the concept is deployed and some worldly property or kind, presumably having to do with water. The problem with narrow content, construed as short-armed functional role or as a function from contexts to wide contents, is that it is not clear how it could involve any such relation. (Kriegel 2008, p. 308)

At this point, however, we seem to butt up against a classic paradox. On the one hand, those theories that appear to capture the relational character of mental content (i.e., causal theories) hold that content is not wholly determined by the intrinsic properties of the brain and, hence, imply that it isn't causally efficacious of behaviour. On the other hand, theories with the potential to account for the causal ef-

ficacy of mental content (i.e., narrow content theories), fail to capture its relational character. A solution to the content causation problem thus requires something that *prima facie* appears impossible: an explanation of the *relational* character of mental content that invokes only the *intrinsic* properties of the brain. Little wonder then that many philosophers despair of ever finding a solution to this puzzle.³

It is reasonable to hazard, however, that one of the main barriers standing in the way of a more productive treatment of the content causation problem is the radically underdeveloped understanding of mental content with which contemporary philosophy operates. In the foregoing quotation, for example, Kriegel characterises the relational character of content in terms of a subject's "cognitive contact" with the external world; yet he readily admits elsewhere that this notion is "not altogether transparent" (Kriegel 2008, p. 305). This is typical of the literature on this topic, which has become accustomed to describing content using the notoriously vague language of *aboutness*. While this language might capture our commonsense intuitions about mental phenomena, its imprecision may prevent us from discerning the lineaments of candidate solutions to the content causation problem.

This last point, at least, gives us the motivation for intruding yet another discussion into this already crowded philosophical space. The foundational conjecture upon which this paper is based is that the apparent insolubility of the content causation problem issues from an impoverished and unenlightening account of the relational character of mental content. Furthermore, this impoverishment is largely a con-

³ Perhaps the best we can do, according to some of these, is to accept that the representational properties of mental phenomena are causally inert, but to argue that there is enough room between explanation and causation for representational properties to be *explanatorily relevant*—despite their inertness (Baker 1993; Block 1989; Fodor 1986, 1989; Heil & Mele 1991; Jackson & Pettit 1990a, 1990b; LePore & Loewer 1989). A more radical response is to opt out of representation-based explanation altogether, as advocated originally by eliminativists (Churchland 1981; Stich 1983), and more recently by anti-representationalists (Brooks 1991). Finally, note that another radical position currently fashionable in philosophy—the extended-mind hypothesis (Menary 2010)—doesn't represent a solution to the content causation problem, since it signally fails to align mental phenomena with the brain-based causation of behaviour.

sequence of the *dyadic* conception of mental representation that has hitherto conditioned most philosophical thinking in this area. By contrast, a minority of philosophers has argued that mental representation is more properly analysed as a *triadic* relation. Triadicity, I will argue, yields a richer and ultimately more illuminating account of the relational character of mental content. Armed with this alternative treatment, we are in a position to assess the content causation problem anew. On the one hand, this novel viewpoint confirms the worry philosophers have expressed that causal theories of mental content are impossible to reconcile with the brain-based causation of behaviour. On the other hand, and much more positively, the triadic conception reveals a path that, from the perspective of content causation at least, looks more promising. The proposal that we travel down this path will undoubtedly face resistance, since it requires us to rehabilitate an approach to mental content that is unpopular in contemporary philosophy. But this approach, I shall conclude, seems unavoidable if we are to explain how mind matters.

2 The triadicity of representation

The bulk of philosophical writing on representation in general and mental representation in particular assumes, either explicitly or implicitly, that representation is a dyadic relation between something that does the *representing* and something that is *represented*. The task for a theory of representation, from this perspective, is to explain the necessary and sufficient conditions under which this dyadic relation obtains (see e.g., Stich 1992). But such a dyadic conception provides very little guidance about the relational character of representational content. All we have to work with is a mysterious action-at-a-distance phenomenon, whereby one part of the world, in virtue of the obtaining of a certain relation, is about another part.

To fill this gap, philosophers have almost invariably modelled their understanding of content on the semantic properties of the elements that compose our natural languages. Given the towering influence of Tarskian truth-conditional

semantics in this field, it is inevitable that the relational character of representational content is usually characterised in terms of *reference* (Kriegel 2008, p. 305). But such an approach, while perhaps appropriate for linguaform representation, sits awkwardly with all manner of the non-linguistic forms of representation with which we are familiar (Haugeland 1991; Fodor 2007; Cummins & Roth 2012). Moreover, it is not obvious we are more enlightened by replacing talk of aboutness with that of reference.

In this context, it is worth observing that over the years a minority of philosophers has expressed dissatisfaction with the dyadic conception of representation. The most salient complaint is that such an approach fails to take into consideration the role that “users” of representation play. The general thought here is that some parts of the world don’t represent other parts solely in virtue of some relationship between them; that the former represent the latter only when they are employed by some system to perform this representational function. According to Dennett, for example, physical entities “are by themselves quite inert as information bearers. [...] They become information-bearers only when given roles in larger systems” (1982, p. 217). Likewise, Millikan has long observed that a biological approach to representation forces one to consider not just the “production” of representations, but also their “consumption” (1984). And, in a similar vein, Bechtel argues that that since whether something acts as a representation is ultimately determined by its function for some user, it follows that there are “three interrelated components in a representation story: what is represented, the representation, and the user of the representation” (1998, p. 299).

This *triadic* conception of representation is not new, of course, since it forms the basis of Charles Sanders Peirce’s theory of semiotics, which was developed in the latter part of the 19th century (Hardwick 1977). Indeed, Peirce’s (sometimes obscure) writings embody one of the most comprehensive analyses of representation in all of philosophical literature. Peirce approached this issue principally by investigating those public forms of representation with which

we are all familiar—words, sentences, paintings, photographs, sculptures, maps, and so forth—but he also sought to apply his triadic analysis to the special case of mental representation. This suggests that Peirce’s writings might be an appropriate point of departure for exploring what the triadic conception entails about the relational character of representational content.

This strategy is very effectively adopted by von Eckardt when, following Peirce’s lead, she analyses representation as a triadic relation involving a “representing vehicle, a represented object, and an interpretation” (von Eckardt 1993, pp. 145-149).⁴ As with dyadic stories, the representing vehicle is the physical object (e.g., a spoken or written word, painting, map, sculpture, etc.) that is about something, and the represented object is the object, property, event, relation, or state of affairs that the vehicle is about. It is the addition of the interpretative relatum that sets the triadic account apart:

A sign [i.e., a representing vehicle] [...] is something which stands to somebody for something in some respect or capacity. It addresses somebody, that is, creates in the mind of that person an equivalent sign, or perhaps a more developed sign. That sign which is created I call the *interpretant* of the first sign. (von Eckardt 1993, p. 145)

Interpretation is thus understood as a cognitive effect in the subject for whom the vehicle operates as a representation. But as von Eckardt observes, not any kind of effect will do. This cognitive effect, presumably implicating the production of *mental* representing vehicles, must bring the subject into some appropriate relationship to the original vehicle’s represented object (von Eckardt 1993, p. 157). Given this constraint, it is natural to interpret this third relatum in terms of the subject’s *thinking* about the object in question. So (non-mental) representation, on the triadic story, is a *functional* kind: it is a process whereby a representing

vehicle triggers a thought (or thoughts) in a subject about a represented object.

There are a couple of significant consequences of the triadicity of representation. The first is that, contrary to a dyadic story, representing vehicles aren’t about anything independent of interpretations. Words, sentences, paintings, photographs, sculptures, maps, and so forth, considered in isolation from the cognitive impact they have on us, don’t represent. This, of course, does some violence to the way that we talk about public representing vehicles—but it is far from catastrophic. The relevant revision is to think of these physical entities as possessing the *capacity* to trigger the necessary cognitive effects in us. The second (and, for our purposes, more important) consequence is that, unlike dyadic accounts in which content is unpacked solely in terms of relations between vehicles and represented objects, the triadic story entails that content is also conditioned by the interpretative relatum. This imposes an additional explanatory requirement on theories of content determination. It is not enough to merely explain how relations between vehicles and objects make it the case that the former are about the latter. These theories must also explain how it is in virtue of these relations that representing vehicles are capable of triggering thoughts in subjects about represented objects.

Once it has been suitably modified for the special, and presumably foundational, case of *mental* representation, the additional explanatory requirement that triadicity imposes on theories of content determination can form the basis of a richer account of the relational character of mental content. Such modification is necessary, of course, because treating the interpretation of mental vehicles solely in terms of a subject thinking about a represented object violates the *naturalism constraint*. This is the requirement that we explain mental representation without recourse to the antecedently representational (see e.g., Cummins 1989, pp. 127–129; Cummins 1996, pp. 3–4; Dretske 1981, p. xi; Fodor 1987, pp. 97–98; Millikan 1984, p. 87; von Eckardt 1993, pp. 234–239).

The relevant modification is fairly obvious, however, and represents a well-trodden path in

⁴ Von Eckardt actually uses the terms “representation bearer”, “representational object”, and “interpretant” to describe the three relata implicated in representation. I prefer the terminology I have used here because it is more consistent with the philosophical literature on mental representation.

philosophy. From the perspective of Peirce's triadic analysis, the role of interpretation is to forge a psychologically efficacious connection between the user of a representing vehicle and the vehicle's object. With public forms of representation it is perfectly acceptable to unpack this in terms of the (non-mental) vehicle activating thoughts directed at the object. But if we allow this story to run a little further it will point us in the right direction for the interpretation of mental vehicles too. Thoughts directed at objects modify our behavioural dispositions towards these same objects. This is why public forms of representation are so useful—they enable us to regulate our behaviour towards selective aspects of the world. But this story can be transported into the brain in order to account for the interpretation of mental representing vehicles. Instead of external vehicles triggering thoughts, and these in turn modifying behavioural dispositions, we simply suppose that mental vehicles have the same cognitive and ultimately behavioural effects. This acts to block the threatened regress since, presumably, it is possible to unpack behavioural dispositions without invoking further mental representation.

We are now in a position to deliver on one of the aims enumerated in the introductory section: that of fashioning a more illuminating account of the relational character of mental content. We saw earlier that Kriegel describes this character in terms of the “cognitive contact” between mental phenomena and the worldly aspects they represent, but admits that this notion isn't particularly transparent. Happily, the triadic analysis of mental representation affords a means of explicating what this cognitive contact consists in. Rather than simply employing the vague language of aboutness, the triadic analysis encourages us to understand the relational character of mental content in terms of the capacity of mental phenomena to regulate the behaviour of subjects towards specific aspects of the world. Cognitive contact is thus a relatively straightforward causal capacity. It is the capacity of cognitive creatures, bestowed by their internal states, to respond selectively to elements of the environment in which they are embedded.

This is where things currently stand. A solution to the content causation problem requires something that *prima facie* appears impossible—namely, an explanation of the relational character of mental content that invokes only the intrinsic properties of the brain. But the paradoxical appearance of content causation, I have suggested, might be a legacy of the dyadic conception of representation that has conditioned philosophical intuitions about content determination, but which provides little guidance about the relational character of mental content. The triadic analysis of representation, I have argued, generates a more enlightening account of this relational character—one pitched in terms of the causal capacities of cognitive creatures to regulate their behaviour towards specific aspects of their environments. From the perspective of this analysis, therefore, a solution to the content causation problem requires a theory of content determination to explain how relations between mental vehicles and their represented objects can endow subjects with the capacity to respond selectively to those very features of the world.

Philosophers seeking to fashion theories of mental content determination over the centuries have famously focused on just two kinds of relations between mental vehicles and their represented objects: “causal” relations and “resemblance” relations (Fodor 1984, pp. 232–233). In the following section I shall engage in an all-too-brief appraisal of the prospects of these two world-mind relations to deliver a solution to the content causation problem.

3 World-mind relations and the content causation problem

Causal theories of mental content determination have dominated philosophy for nearly half a century. They hold that representing vehicles are contentful in virtue of being (actually, nomologically, or historically) caused by their represented objects (Devitt & Sterelny 1987; Fodor 1984, 1987, 1990; Stampe 1977, 1986). Perhaps the most well-known causal theory in all of the literature has been developed, through a number of iterations, by Dretske (1981, 1988, 1995).

What makes Dretske's account particularly apposite in the current context is that it has been fashioned, at least in its later iterations, to address explicitly the account of content intruded by the triadic analysis of mental representation (though Dretske doesn't use this terminology). At one point in his discussion, for example, Dretske states that he approves of Armstrong's (1973) description of beliefs as "maps by which we steer", and goes on to observe that "beliefs are representational structures that acquire their meaning, their maplike quality, by actually *using* the information it is their function to carry in steering the system of which they are part" (Dretske 1988, p. 81). This, for Dretske, motivates the very desideratum we extracted from the triadic analysis in the last section:

It will not be enough merely to have a C [inner state of some cognitive system] that indicates F [i.e., causally covaries with some external condition] cause M [some observable behaviour]. What needs to be done [...] is to show how the existence of one relationship, the relationship underlying C's semantic character, can explain the existence of another relationship, the causal relationship (between C and M) comprising the behaviour in question. (1988, p. 84)

Dretske's response to this problem, famously, is to appeal to teleology. It is only when an inner state, which causally covaries with some bit of the external environment, is "recruited" by the cognitive system (either by an evolutionary design process or through individual development) to cause appropriate behaviour, that the state acquires the *function* of indicating that part of the environment, and thereby comes to *represent* it (Dretske 1988, pp. 84–89).

On the face of it, Dretske's theory seems to represent a promising solution to the content causation problem. From the perspective of the triadic analysis, a solution to this problem requires an explanation of how certain relations between mental vehicles and their objects can dispose cognitive subjects to behave selectively towards those represented objects. Dretske's el-

egant proposal is that reliable causal relations between inner states and environmental conditions (i.e., when the latter reliably cause the former to be tokened) can endow cognitive systems with these dispositions when the former states are conscripted by design processes to cause behaviour that is in some way relevant to the latter conditions. When this happens, the inner states are elevated to the status of representing vehicles, and their subsequent activity in bringing about behaviour directed towards their represented objects are examples of content causation.

Unfortunately, a closer inspection of Dretske's suggestion reveals a fundamental flaw. Contrary to what he contends, the relations at the core of his proposal are powerless to explain the required behavioural dispositions. Rather than describing this problem in the abstract, let me illustrate it using one of Dretske's favourite examples of a very simple representation-using system:

A drop in room temperature causes a bi-metallic strip in [a thermostat] to bend. Depending on the position of an adjustable contact, the bending strip eventually closes an electrical circuit. Current flows to the furnace and ignition occurs. The thermostat's behaviour, its turning the furnace on, is the bringing about of furnace ignition by events occurring in the thermostat—in this case [...] the closure of a switch by the movement of a temperature-sensitive strip [...].

The bi-metallic string is given a job to do, made part of an electrical switch for the furnace, because of what it indicates about room temperature. Since this is so, it thereby acquires the function of indicating what the temperature is [...]. We can speak of [...] representation here. (Dretske 1988, pp. 86–87)

There is a subtle sleight of hand at work here, however. It is Dretske's contention that the bi-metallic strip is recruited (by the manufacturer) to play a causal role in the thermostat because

of what it indicates about ambient temperature. But that's not the full story. Bi-metallic strips have an *additional* property that appeals to the manufacturers of thermostats: their degree of curvature corresponds in an orderly fashion with ambient air temperature, such that it can be configured to complete a circuit when the temperature drops to a pre-set level.

In Dretske's thermostat example, therefore, there are two distinct relations between representing vehicles and represented objects: a systematic correspondence relation (wherein variations in ambient air temperature are mirrored by orderly variations in the bi-metallic strip's shape) and an indication relation (wherein variations in ambient air temperature *cause* variations in the bi-metallic strip).⁵ These two relations are not independent of one another, of course, as the former is mediated by the latter. But we can still consider which of these relations is doing the work, insofar as the capacity for the thermostat to control the behaviour of the furnace is concerned. And here the answer is clear: it is the fact that the curvature of the bi-metallic strip systematically mirrors the temperature, and not the causal covariation per se, that explains its capacity to operate the furnace in an appropriate manner. Consider the counterfactuals: curvature correspondence without causal covariation (e.g., where a mere correlation exists) would still generate the appropriate behaviour, but causal covariation without curvature correspondence (e.g., where the bi-metallic strip heats up but maintains its shape) wouldn't. The important point is that while the causal relation plays an important role in mediating the correspondence relation, it is the latter, not the former, that explains

the thermostat's capacity to bring about the desired behaviour.⁶

So Dretske's own example fails to satisfy the desideratum that he set for himself: the obtaining of a reliable causal connection between ambient air temperature and the bi-metallic strip doesn't explain the thermostat's capacity to control the behaviour of the furnace. Moreover, this example illustrates a fundamental problem with *all* causal theories of mental content determination: there is a fatal disconnect between world-mind causal relations, on the one hand, and the mind's behavioural dispositions on the other. This disconnect exists because any (actual, nomological, or historical) causal relations that might exist between external conditions and inner vehicles do not explain, in their own right, how a cognitive system inherits the capacity of behaving sensitively to the former. Whether cognitive systems have this capacity is determined by the properties of their inner vehicles in concert with their organizational, architectural, and motoric properties. And while external conditions can cause tokenings of and alterations to inner vehicles, the mere obtaining of such causal relations can't explain how the tokened or altered vehicles are capable of interacting with these multifarious systemic properties such that they bestow the appropriate behavioural dispositions.⁷ This is why manufacturers are very choosy about the materials from which they construct thermo-

⁵ Dretske scholars will cry foul at this point, of course. This is because Dretske claims that while indication is mostly founded on causal relations, it need not be. Indeed, he goes as far as to suggest that indication obtains whenever there is a non-coincidental covariation between vehicle and object (Dretske 1988, pp. 56–57). But this characterisation of indication transforms Dretske's proposal into something close to a resemblance theory (the approach to be examined in the next section), since it privileges systematic correspondence relations over causal relations. Consequently, insofar as Dretske's position is to be understood as a causal theory of content determination (as is widely assumed in the literature), it is essential that indication is interpreted as a relation of causal covariation. I adopt this interpretation in what follows.

⁶ One would expect to find causal relations mediating systematic correspondence relations between the representing vehicles of biological systems and aspects of the world. But, as Dretske is well aware, this is not always the case. Nature will make do with what works, and some kind of systematic correspondence in the absence of causal commerce will do just as well. This can be illustrated by another of Dretske's favourite examples: the evolutionary recruitment of magnetosomes in anaerobic bacteria to steer them towards deoxygenated water (1986). According to Dretske, evolutionary forces operating on these bacteria have selected magnetosomes because they are indicators of anaerobic water capable of influencing the direction in which the bacteria swim. But as Millikan has pointed out, the connection between the orientation of magnetosomes and anaerobic water is merely correlational, not causal (2004, Ch. 3). Magnetosomes indicate and steer northern hemisphere anaerobic bacteria in the direction of magnetic north, which results in these bacteria swimming into deeper (and hence deoxygenated) water. But there is no causal connection between magnetic north and deoxygenated water. In this case, therefore, magnetosomes have been selected because their alignment systematically corresponds with the direction of anaerobic water, not in virtue of any causal covariation between them.

⁷ Cummins reaches a similar conclusion, though via a somewhat different route (1996, p. 74).

stats. Engineering a causal covariation relation between ambient air temperature and the innards of a thermostat is easy; engineering these innards such that they possess the requisite causal capacities is a great deal harder.

Ultimately, therefore, Dretske's ingenious attempt to solve the content causation problem doesn't succeed. Dretske holds that the internal states of cognitive systems are elevated to representing roles when they are recruited by design processes to regulate behaviour towards the external conditions they indicate. He takes this to be a case of genuine content causation because he thinks that the causal relations between represented objects and representing vehicles can explain the causal activity in which the vehicles subsequently engage. But Dretske has over-estimated the explanatory power of world–mind causal relations. And he has done so because he has illicitly smuggled into his story a quite distinct form of content determination—one that exploits systematic correspondence relations between representing vehicles and their represented objects. Such systematic correspondences are, of course, a species of resemblance relation. The failure of Dretske's proposal is thus instructive, since it suggests that this alternative world–mind relation offers some prospect of a solution to the content causation problem.

Resemblance theories of content determination hold that representing vehicles are contentful in virtue of resembling their represented objects. The most obvious and straightforward application of this idea can be found in many public forms of representation, from photographs and paintings to sculptures and maps. But what is most significant about this approach for our purposes is that when vehicles resemble their objects, the former actually *replicate* the latter in some way, either by reproducing their properties or their relational organisation (more about which in the next section). And this affords a relatively straightforward way of explaining how a physical device, in virtue of incorporating vehicles that bear resemblance relations to the world, acquires a capacity to behave selectively towards particular elements of the environment. The thermostat's bi-

metallic strip reproduces—in the variations in its degree of curvature—the diachronic pattern of magnitude relations between ambient air temperature. Once this bi-metallic strip is incorporated into the thermostat, therefore, this device has a set of internal vehicles that dynamically replicates the external temperature. It is then simply a matter of rigging the innards of the thermostat so that its operation of the furnace is regulated by these internalised surrogates (Swoyer 1991).

Dretske is correct to judge this an example of content causation. It is a case in which the exploitation of a relation between environmental conditions and inner vehicles explains how the latter are capable of modifying a device's behavioural dispositions towards particular aspects of the world. But what is seldom acknowledged about this much-used example is that it demonstrates the causal efficacy of content fixed by resemblance. Despite this virtue, resemblance theories of mental content determination are unfashionable in contemporary philosophy, largely because they are widely thought to suffer from a number of fatal flaws. Before we end, therefore, it would be wise to engage in a degree of resemblance rehabilitation. This turns out to be easier than one might expect, however, once we adopt the perspective of the triadic conception of representation.

4 Rehabilitating resemblance

Despite the widespread assumption that they are fatally flawed, it's hard to find a sustained discussion of the problems associated with resemblance theories of content determination. Instead, one finds scattered somewhat haphazardly across the literature brief allusions to the same five objections. The canonical rendering of three of these can be found in the opening paragraphs of Nelson Goodman's *Languages of Art*:

The most naive view of representation might perhaps be put somewhat like this: "A represents B if and only if A appreciably resembles B", or "A represents B to the extent that A resembles B". Vestiges of this view, with assorted refinements, per-

sist in most writing on representation. Yet more error could hardly be compressed into so short a formula.

Some of the faults are obvious enough. An object resembles itself to the maximum degree but rarely represents itself; resemblance, unlike representation, is reflexive. Again, unlike representation, resemblance is symmetric: B is as much like A as A is like B, but while a painting may represent the Duke of Wellington, the Duke doesn't represent the painting. Furthermore, in many cases neither one of a pair of very like objects represents the other; none of the automobiles off an assembly line is a picture of any of the rest; and a man is not normally a representation of another man, even his twin brother. Plainly, resemblance in any degree is no sufficient condition for representation. (1969, pp. 3–4)

In short, representation can't be based on resemblance, since the latter is *reflexive* (where the former isn't), *symmetric* (where the former isn't), and *insufficient* (all manner of objects resemble others without representing them). But however influential these three objections might be when applied to a dyadic analysis of representation, they lose all force in the context of a triadic conception. This conception agrees with Goodman that relations between vehicles and their represented objects are insufficient to confer representational status. A representing vehicle must also undergo interpretation, either by triggering thoughts in a cognitive subject or by modifying the subject's behavioural dispositions. And it is this process of interpretation, according to a resemblance theory, that also enforces the non-reflexivity and asymmetry of representation.

A fourth objection is that resemblance theories of mental content are incompatible with our commitment to physicalism:

If mental representations are physical things, and if representation is grounded in [resemblance], then there must be phys-

ical things in the brain that are similar to (i.e., share properties with) the things they represent. This problem could be kept at bay only so long as mind-stuff was conceived as nonphysical. The idea that we could get redness and sphericity in the mind loses its plausibility if this means we have to get it in the brain. (Cummins 1989, p. 31)

But this objection is easily deflected once a proper understanding of the different forms of resemblance is in place. The most straightforward kind of resemblance—the kind that Cummins in the above quotation has in mind, for example—involves the sharing of one or more properties. This relationship can be termed *first-order resemblance*.⁸ It is this kind of resemblance that grounds the content of many public forms of representation, such as paintings, sculptures, and scale models. As Cummins points out, however, first-order resemblance is clearly unsuitable as a ground of mental content, since it is incompatible with what we know about the brain.

There is, nonetheless, a more abstract species of resemblance available. The requirement that representing vehicles share properties with their represented objects can be relaxed in favour of one in which the *relations* among a system of vehicles mirror the *relations* among their objects. This relation-preserving mapping between two systems can be called *second-order resemblance*.⁹ And while it is extremely unlikely

⁸ I am here adapting terminology used by Shepard & Chipman (1970).

⁹ To be more precise, suppose $S_V = (V, \nu)$ is a system comprising a set V of objects, and a set ν of relations defined on the members of V . The objects in V may be conceptual or concrete; the relations in ν may be spatial, causal, structural, or inferential, and so on. For example, V might be a set of features on a map, with various geometric and part-whole relations defined on them. Or V might be set of well formed formulae in first-order logic falling under relations such as identity and consistency. There is a *second-order resemblance* between two systems $S_V = (V, \nu)$ and $S_O = (O, o)$ if, for at least *some* objects in V and *some* relations in ν , there is a one-to-one mapping from V to O and a one-to-one mapping from ν to o such that when a relation in ν holds of objects in V , the corresponding relation in o holds of the corresponding objects in O . In other words, the two systems resemble each other with regard to their abstract relational organisation. As already stressed, resemblance of this kind is independent of first-order resemblance, in the sense that two systems can resemble each other at second-order without sharing properties. Second-order resemblance comes in weaker and stronger forms. As defined it is relatively weak, but if we insist on a mapping that takes

that first-order resemblance is the general ground of mental content (given what we know about the brain), the same does not apply to second-order resemblance. Two systems can share a pattern of relations *without* sharing the physical properties upon which those relations depend. Second-order resemblance is actually a very abstract relationship: essentially nothing about the physical form of a system of representing vehicles is implied by the fact that it resembles a set of represented objects at second-order. Contrary to the fourth objection, therefore, a theory of mental content determination that exploits second-order resemblance is compatible with physicalism.¹⁰

However, this emphasis on second-order resemblance, at least in the eyes of many theorists, takes this approach to content determination out of the frying pan and into the fire. This is because the highly abstract nature of second-order resemblance invites the charge that it entails a massive and intractable indeterminacy of mental content. And it is this fifth objection, perhaps more than any other, that accounts for the current unpopularity of resemblance theories (Sprevak 2011).

The problem here can be illustrated by returning to Dretske's thermostat. The world-mind relation that does all the heavy lifting here constitutes a second-order resemblance: the relations among the representing vehicles (the set of bi-metallic strip curvatures) systematically mirror the relations among the representing objects (the set of ambient air temperatures).¹¹ The worry embodied in the fifth objection, how-

ever, is that this same set of representing vehicles will second-order resemble not just the temperature surrounding the thermostat but any set of objects, regardless of their nature and location, that shares its relational organisation. This fact is entailed by the abstract nature of second-order resemblance. And this is a problem, of course, because it suggests that second-order resemblance is incapable of delivering determinate content. The most we can say about the thermostat's bi-metallic strip is that its curvature represents that potentially large and motley collection of objects with which it systematically corresponds. And this would seem to be a long way from saying it represents the temperature of the ambient air.

Fortunately, the triadic analysis again offers a way to surmount this difficulty. On this account, representations aren't manufactured solely from relations between vehicles and the objects they represent. Rather, the process of interpretation must also be thrown into the mix. We've seen that interpretation is discharged ultimately in terms of modifications to a system's behavioural dispositions. But not any old modifications will do—representing vehicles must modify the system's dispositions towards their represented objects. Consequently, interpretation plays an important content-limiting role. Specifically, a system's behavioural dispositions will anchor its representing vehicles to particular represented domains. Once a domain is secured in this way, second-order resemblance relations determine the content of the individual vehicles. In the case of the thermostat, for example, the behavioural dispositions of the system restrict the represented domain to ambient air temperature, and the second-order resemblance relations determine what temperature each vehicle represents.

¹¹ Notice that in this case, the second-order resemblance is sustained *structural* relations among the set of representing vehicles (i.e., the set of bi-metallic strip curvatures). This is an example of what Palmer (1978) calls *natural isomorphism*, since the second-order resemblance relations are sustained by constraints *inherent* in the vehicles, rather than being imposed *extrinsically*. Elsewhere I have used the term *structural resemblance* to describe this kind of second-order relationship and to distinguish it from *functional resemblance*, where the second-order resemblance relations are sustained by *causal* relations among the vehicles—see O'Brien & Opie (2004).

every element of V onto some element of O , and, in addition, preserves *all* the relations defined on V , then we get a strong form of resemblance known as a *homomorphism*. Stronger still is an *isomorphism*, which is a one-to-one relation-preserving mapping such that every element of V corresponds to some element of O , and every element of O corresponds to some element of V . When two systems are isomorphic their relational organisation is identical. In the literature on second-order resemblance the focus is often placed on isomorphism (see e.g., Cummins 1996, pp. 85–111), but where representation is concerned, the kind of correspondence between systems that is likely to be relevant will generally be weaker than isomorphism. For a much fuller discussion of second-order resemblance, see O'Brien & Opie (2004).

¹⁰ Two early theorists who sought to apply second-order resemblance to mental representation are Palmer (1978) and Shepard (Shepard & Chipman 1970; Shepard & Metzler 1971). More recently, Blachowicz (1997), Cummins (1996), Gardenfors (1996), O'Brien & O'Brien (1999), O'Brien & Opie (2004), and Swoyer (1991), have all defended second-order resemblance theories.

5 Conclusion: How mind matters

It is time to take stock. We began with three commonplace theses about mental phenomena and their physical realization in the brain that together generate a profound puzzle about mental causation. This is the content causation problem: that of explaining how the specifically representational properties of mental phenomena can be causally efficacious of behaviour. This problem has an air of insolubility about it because it appears to require something impossible: an explanation of the relational character of mental content that invokes only the intrinsic properties of the brain. It has been the foundational conjecture of this discussion, however, that this despair issues from the impoverished understanding of content that attends the dyadic analysis of mental representation, and that once we adopt the perspective of the triadic conception our view of the content causation problem is transformed.

The insight offered by triadicity is that the relational character of mental content is to be discharged ultimately in terms of our behavioural dispositions towards features of the world. This offers a way forward with the content causation problem because it suggests that, rather than seeking to explain some kind of mysterious action-at-a-distance, the task for a theory of content determination is to explain how the obtaining of world-mind relations can dispose cognitive systems to respond selectively to certain elements of their embedding environments.

According to most naturalistically-inclined philosophers, there are just two candidate mind-world relations available: causal relations and resemblance relations. Causal theories of content determination dominate the contemporary landscape, but our analysis confirms what many have suspected—namely, that causal theories offer no prospect of a solution to the content causation problem. The reason for this, however, is not because such theories appeal to relations that incorporate factors beyond the brain. All theories of mental representation, in their efforts to explain the relational character of mental content, are forced to invoke world-

mind relations of some kind. The problem with causal theories, at least from the triadic perspective, is the disconnect between world-mind causal relations and a system's behavioural dispositions. The obtaining of causal relations between external conditions and inner vehicles cannot explain how the latter endow systems with the capacity to respond in a discriminating fashion towards the former.

This leaves us with resemblance relations. The problem here is that resemblance theories of content determination have for many years been deeply unpopular in philosophy. But this is another point where the triadic conception of representation pays rich dividends. Most of the problems associated with resemblance theories don't look so severe when viewed from the triadic perspective. This is encouraging, because resemblance does offer some prospect of a solution to the content causation problem. The key here is that the mere obtaining of the resemblance relation entails that representing vehicles replicate their represented objects. This ensures that the former have properties that can be exploited to shape the behavioural dispositions of cognitive systems towards the latter.

Consequently, if the line of argument presented here is on the right track, then resemblance theories of mental content determination must be rehabilitated and subjected to scrutiny and development. It goes without saying that there are a great number of significant hurdles yet to be overcome. I have focused, for instance, on just one very simple example of a representation-using system. There remains, accordingly, a large question mark over whether the resemblance solution to the problem of content causation scales up to more sophisticated cognitive creatures, let alone to the immense complexities of our own mental phenomena. But we have to start somewhere. And as things currently stand, resemblance theories appear to be obligatory, since they alone offer some prospect for explaining how mind matters.

References

- Adams, F. & Aizawa, K. (2010). Causal theories of mental content.
<http://plato.stanford.edu/entries/content-causal/>
- Armstrong, D. (1973). *Belief, truth and knowledge*. Cambridge, UK: Cambridge University Press.
- Baker, L. R. (1993). Metaphysics and mental causation. In J. Heil & A. Mele (Eds.) *Mental Causation* (pp. 75-96). Oxford, UK: Oxford University Press.
- Bechtel, W. (1998). Representations and cognitive explanations: Assessing the dynamicist's challenge in cognitive science. *Cognitive Science*, 22 (3), 295-318.
[10.1207/s15516709cog2203_2](https://doi.org/10.1207/s15516709cog2203_2)
- Blachowicz, J. (1997). Analog representation beyond mental imagery. *Journal of Philosophy*, 94 (2), 55-84.
- Block, N. (1986). Advertisement for a semantics for psychology. *Midwest Studies in Philosophy*, 10 (1), 615-678. [10.1111/j.1475-4975.1987.tb00558.x](https://doi.org/10.1111/j.1475-4975.1987.tb00558.x)
- (1989). Can the mind change the world? In G. Boolos (Ed.) *Meaning and method: Essays in honor of Hilary Putnam* (pp. 137-170). Cambridge, UK: Cambridge University Press.
- Brooks, R. (1991). Intelligence without representation. *Artificial Intelligence*, 47 (1-3), 139-159.
[10.1016/0004-3702\(91\)90053-M](https://doi.org/10.1016/0004-3702(91)90053-M)
- Burge, T. (1979). Individualism and the mental. *Midwest Studies in Philosophy*, 4 (1), 73-122.
[10.1111/j.1475-4975.1979.tb00374.x](https://doi.org/10.1111/j.1475-4975.1979.tb00374.x)
- (1986). Individualism and psychology. *Philosophical Review*, 95 (1), 3-45. [10.1007/978-94-009-2649-3_3](https://doi.org/10.1007/978-94-009-2649-3_3)
- Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy*, 78 (2), 67-90. [10.2307/2025900](https://doi.org/10.2307/2025900)
- Crane, T. (1995). The mental causation debate. *Proceedings of the Aristotelian Society*, 69, 211-236.
- Cummins, R. C. (1989). *Meaning and mental representation*. Cambridge, MA: MIT Press.
- (1996). *Representations, targets, and attitudes*. Cambridge, MA: MIT Press.
- Cummins, R. C. & Roth, M. (2012). Meaning and content in cognitive science. In R. Schantz (Ed.) *Prospects for meaning* (pp. 365-382). Berlin, GER: de Gruyter.
- Dennett, D. (1978). *Brainstorms*. Cambridge, MA: MIT Press.
- (1982). Styles of mental representation. *Proceedings of the Aristotelian Society*, 83, 213-226.
- (1987). *The intentional stance*. Cambridge, MA: MIT Press.
- Devitt, M. & Sterelny, K. (1987). *Language and reality*. Cambridge, MA: MIT Press.
- Dretske, F. (1981). *Knowledge and the flow of information*. Oxford, UK: Clarendon.
- (1986). Misrepresentation. In R. Bogdan (Ed.) *Belief: Form, content, and function* (pp. 17-36). Oxford, UK: Oxford University Press.
- (1988). *Explaining behavior: Reasons in a world of causes*. Cambridge, MA: MIT Press.
- (1995). *Naturalizing the mind*. Cambridge, MA: MIT Press.
- Fodor, J. (1984). Semantics, Wisconsin style. *Synthese*, 59 (3), 231-250. [10.1007/BF00869335](https://doi.org/10.1007/BF00869335)
- (1986). Banish discontent. In J. Butterfield (Ed.) *Language, mind, and logic* (pp. 1-24). Cambridge, UK: Cambridge University Press.
- (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. Cambridge, MA: MIT Press.
- (1989). Making mind matter more. *Philosophical Topics*, 17 (1), 59-79. [10.5840/philtopics198917112](https://doi.org/10.5840/philtopics198917112)
- (1990). *A theory of content and other essays*. Cambridge, MA: MIT Press.
- (2007). The revenge of the given. In B. McLaughlin & J. Cohen (Eds.) *Contemporary debates in the philosophy of mind* (pp. 105-116). Malden, MA: Blackwell.
- Gärdenfors, P. (1996). Mental representation, conceptual spaces and metaphors. *Synthese*, 106 (1), 21-47.
[10.1007/BF00413612](https://doi.org/10.1007/BF00413612)
- Goodman, N. (1969). *Languages of art*. London, UK: Oxford University Press.
- Hardwick, C. (Ed.) (1977). *Semiotics and signification: The correspondence between Charles S. Peirce and Victoria Lady Welby*. Bloomington, IN: Indiana University Press.
- Haugeland, J. (1991). Representational genera. In W. Ramsey, S. Stich & D. Rumelhart (Eds.) *Philosophy and connectionist theory* (pp. 171-206). Hillsdale, NJ: Lawrence Erlbaum.
- Heil, J. & Mele, A. (1991). Mental causes. *American Philosophical Quarterly*, 28 (1), 61-71.
- Hohwy, J. (Ed.) (2008). *Being reduced: New essays on reduction, explanation and causation*. New York, NY: Oxford University Press.
- Jackson, F. & Pettit, P. (1990a). Causation and the philosophy of mind. *Philosophy and Phenomenological Research Supplement*, 50, 195-214.
- (1990b). Program explanation: A general perspective. *Analysis*, 50 (2), 107-117.
- Kim, J. (1992). Multiple realization and the metaphysics

- of reduction. *Philosophy and Phenomenological Research*, 52 (1), 1-26. [10.1017/CBO9780511625220.017](https://doi.org/10.1017/CBO9780511625220.017)
- (2000). *Mind in a physical world*. Cambridge, MA: MIT Press.
- (2005). *Physicalism, or something near enough*. New York, NY: Princeton University Press.
- (2006). *Philosophy of mind*. Boulder, CO: Westview Press.
- Kriegel, U. (2008). Real narrow content. *Mind and Language*, 23 (3), 304-328. [10.1111/j.1468-0017.2008.00345.x](https://doi.org/10.1111/j.1468-0017.2008.00345.x)
- LePore, E. & Loewer, B. (1989). More on making mind matter. *Philosophical Topics*, 17 (1), 175-191. [10.5840/philtopics198917117](https://doi.org/10.5840/philtopics198917117)
- Loar, B. (1981). *Mind and meaning*. Cambridge, UK: Cambridge University Press.
- (1982). Conceptual role and truth conditions. *Notre Dame Journal of Formal Logic*, 23 (3), 272-283. [10.1305/ndjfl/1093870086](https://doi.org/10.1305/ndjfl/1093870086)
- Menary, R. (Ed.) (2010). *The extended mind*. Cambridge, MA: MIT Press.
- Millikan, R. G. (1984). *Language, thought and other biological categories*. Cambridge, MA: MIT Press.
- (2004). *Varieties of meaning*. Cambridge, MA: MIT Press.
- O'Brien, G. & O'Brien, J. (1999). Putting content into a vehicle theory of consciousness. *Behavioral and Brain Sciences*, 22 (1), 175-196.
- O'Brien, G. & Opie, J. (2004). Notes towards a structuralist theory of mental representation. In P. S. Clapin & P. Slezak (Eds.) *Representation in mind: New approaches to mental representation*. Amsterdam, NL: Elsevier.
- Palmer, S. (1978). Fundamental aspects of cognitive representation. In E. Rosch & B. Lloyd (Eds.) *Cognition and categorization* (pp. 259-303). Hillsdale, NJ: Lawrence Erlbaum.
- Putnam, H. (1975). The meaning of 'meaning'. In H. Putnam (Ed.) *Mind, language, and reality* (pp. 131-193). Cambridge, UK: Cambridge University Press.
- Shepard, R. & Chipman, S. (1970). Second-order isomorphism of internal representations: Shapes of states. *Cognitive Psychology*, 1 (1), 1-17. [10.1016/0010-0285\(70\)90002-2](https://doi.org/10.1016/0010-0285(70)90002-2)
- Shepard, R. & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, 171 (3972), 701-703.
- Sprevak, M. (2011). Review of William H. Ramsey's 'Representation Reconsidered'. *British Journal for the Philosophy of Science*, 62 (3), 669-675.
- Stampe, D. (1977). Towards a causal theory of linguistic representation. *Midwest Studies in Philosophy*, 2 (1), 42-63. [10.1111/j.1475-4975.1977.tb00027.x](https://doi.org/10.1111/j.1475-4975.1977.tb00027.x)
- (1986). Verificationism and a causal account of meaning. *Synthese*, 69 (1), 107-137. [10.1007/BF01988289](https://doi.org/10.1007/BF01988289)
- Stich, P. (1983). *From folk psychology to cognitive science*. Cambridge, MA: MIT Press.
- (1992). What is a theory of mental representation? *Mind*, 101 (402), 243-261. [10.1093/mind/101.402.243](https://doi.org/10.1093/mind/101.402.243)
- Swoyer, C. (1991). Structural representation and surrogate reasoning. *Synthese*, 87 (3), 449-508. [10.1007/BF00499820](https://doi.org/10.1007/BF00499820)
- von Eckardt, B. (1993). *What is cognitive science?* Cambridge, MA: MIT Press.
- Wilson, R. (1994). Wide computationalism. *Mind*, 103 (411), 351-372. [10.1093/mind/103.411.351](https://doi.org/10.1093/mind/103.411.351)

Does Resemblance Really Matter?

A Commentary on Gerard O'Brien

Anne-Kathrin Koch

In this commentary on Gerard O'Brien's "How does mind matter?—Solving the content causation problem", I will investigate the notion of *representational content* presented in the latter. With this notion, O'Brien aims at giving an explanation of how mind matters in physicalist terms. His argumentation is motivated by, and supposedly directed towards, a problem he calls *the content causation problem*. Regarding this, I am most interested in reconstructing how his account relates to the presuppositions that make this problem so pressing in philosophical enquiry. O'Brien provides a very interesting answer to the question of "why mental content matters", as motivated by the content causation problem. In particular, I will try to show that by making use of the notion of dispositions, it provides an interesting way of avoiding the presupposition that understanding content causation always requires the reduction of individual relational properties to individual intrinsic properties—probably because it is presupposed that such a reduction is impossible.

Keywords

Dispositions | Mental causation | Mental representation | Reduction

Commentator

Anne-Kathrin Koch

anne-kathrin.koch@gmx.de

Johannes Gutenberg-Universität
Mainz, Germany

Target Author

Gerard O'Brien

gerard.obrien@adelaide.edu.au

University of Adelaide
Adelaide, Australia

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

Gerard O'Brien's paper "How does mind matter?—Solving the content causation problem" ([this collection](#)) is situated at the border of philosophy and cognitive science. The subject matter, as announced by the title, is the causal efficacy of mental content, especially of representational content. O'Brien approaches this subject in three argumentative steps: first he introduces us to a problem called the *content causation problem*, then he proposes a conception of representation that he calls the *triadic conception of representation*, and, third, he enriches this concept by proposing a second-order similarity theory of content determination.

In this commentary I will try to reconstruct how these three points relate to each other, focusing in particular on the role that the content causation problem plays in the other two argumentative steps. O'Brien's theory of representational content and its causal efficacy is doubtlessly interesting even when considered in isolation, as I will briefly outline in section 2. In section 3, I will try to show that the content causation problem demands us to be more specific than when just investigating the question of how mind matters. In section 4, I will try to show how O'Brien's account of the relational character of mental content, which is at the core

of his argumentation, gains its philosophical relevance from implicit assumptions that form the conceptual foundations of the content causation problem as here formulated. In an attempt to assess whether his account must really be regarded as *solving* the content causation problem, I will highlight in section 5 how important it is to be specific about what we really mean if we suppose that representation is somehow *relational* in character.

2 How mind might matter

In his paper in [this collection](#), Gerard O'Brien confronts the task of "explaining how mind matters" (p. 12). He does so, because he—rightly, I believe—identifies the fact "that our minds matter—that our beliefs and desires, and our perceptions and thoughts ultimately have a causal impact on our behaviour" as a ubiquitous and well-accepted, but unexplained phenomenon (p. 1).

O'Brien's investigation is motivated by the following question: "[h]ow can mental phenomena be causally efficacious of behaviour in virtue of their representational contents if those contents are not determined by intrinsic properties of the brain?" (O'Brien [this collection](#), p. 2) He calls this question the "content causation problem" (*ibid.*, p. 2). This specific way of approaching the matter of mental causation is set in the context of "three widely accepted theses about mental phenomena and their physical realization in the brain" (*ibid.*, p. 2): (i) the supposed causal efficacy of mental phenomena is grounded in their representational contents, which, (ii) are taken to be *relational* properties of those phenomena (*ibid.*, p. 2–3); and (iii) the results of neuroscience, which already provide us with an explanation of how behaviour is caused, only make use of the brain and its *intrinsic* properties in their explanation (*ibid.*, p. 2). Hence, there is a need for an explanation of behaviour being caused by mental phenomena in virtue of their relational properties, and this explanation cannot easily make use of the explanation of the causation of behaviour that has already been provided by contemporary

neuroscience. At first, it looks as if this shortfall is exactly what O'Brien is addressing.

Philosophical mainstream accounts of representation, O'Brien reminds us, are built on an understanding of representation as a two-place relation. Representational content is thus described in terms of *aboutness* and/or *reference*. O'Brien, however, advises us to abandon the traditional understanding of the notion of representation, i.e., the idea that representation is a two-place relation and adequately phrased in terms of one thing being *about* another (O'Brien [this collection](#), p. 3–4). Instead, he proposes a triadic conception of representation, making representation a three-place relation between a represented object, a representing vehicle, and an interpretation (*ibid.*, p. 5).

In the triadic picture, interpretation is "a cognitive effect [of the object] in the subject", thereby establishing a relationship between this subject and the represented object (O'Brien [this collection](#), p. 5). The ingenious move here, of course, is that interpretation is explained in *causal* vocabulary. We should think of interpretation as "presumably implicating the production of mental representing vehicles" possessing new properties, which should in turn be thought of as "bring[ing] the subject into some appropriate relationship to the original vehicle's represented object" (*ibid.*, p. 5), i.e., "modifying [the subject's] behavioural dispositions" (*ibid.*, p. 6). At first, when O'Brien further describes those vehicles as "hav[ing] [...] cognitive and ultimately behavioural effects" (*ibid.*, p. 6), it isn't clear exactly which category we are dealing with. I take the relata of the causation relation to be events, but understand talk of vehicles to be talk about objects.¹ I suggest that we understand the vehicles as modifying the system's behavioural dispositions insofar as, once produced, they have certain properties that are directly and specifically relevant for a causal process to take place (given that some sort of stimulus initiates the causal process). If we adopted a view of dispositions as second-order properties, i.e., the property of having certain properties that

¹ If we want to avoid reification of the "vehicles", we might look at it them as time-slices in a complex, internal chain of events, i.e., of dynamic inner processes modifying a subject's global dispositions.

can be causally relevant (cf. Choi & Fara 2014), this would allow us to think of the vehicles as modifying global dispositions of a system as a whole, in the sense of *providing new ones*—that is, by themselves having novel dispositional properties. This way, we can analyze the obtaining of the representation relation as a specific causal process having taken place: the first step of that process is the triggering of the cognitive effect by the representandum (the first relatum); the second relatum is the event of interpretation itself; and the third relatum is the new vehicle produced during the event of interpretation that provides the subject with a new behavioural disposition towards the representandum. The representational character of mental content, in this picture, just rests on what we call “content” resulting from the multi-layer causal process described above.²

O’Brien’s triadic account of representation describes the obtaining of the representation relation not in terms of our everyday intuitions about representation, but in terms of the job it is supposedly doing for us: bridging the gap between whatever is going on in the sphere of “the mental” and the external world by alluding to the causal chain that unites the two. It is thus understandable why the dyadic conception might be accused of hiding behind terms like “aboutness” or “reference”: saying that something mental is about something external is just saying that there is a gap being bridged. Saying that something external sets a three-step causal chain in motion with the result that a subject has undergone a specific change in her behavioural dispositions seems much closer to saying what the bridge is made of.

Yet we should still dispose of the vague language of “specific change in her behavioural

dispositions”. What exactly makes this change specific? It was called “specific” because it selectively relates back to the object that set the causal chain in motion in the first place and which we would like to keep calling “the represented object.”³ But how can the change in a subject’s behavioural dispositions make them pick out the exact same object from which this change originates?⁴ The answer to this lies in the theory of content determination.

When holding that the representing vehicle brings about a change in the subject’s behavioural dispositions, causal theories of content determination are supposedly to be abandoned because of a “disconnect between world–mind causal relations and a system’s behavioural dispositions” (O’Brien this collection, p. 12). An appropriate theory of content determination—so says a desideratum that we gain from the results of the triadic analysis of representation—must “explain how [inner vehicles] endow systems with the capacity to respond in a discriminating fashion towards [external conditions]” (ibid., p. 12). For fulfillment of this criterion, O’Brien turns to *resemblance theories* of content determination, which “hold that representing vehicles are contentful in virtue of resembling their represented objects” (this collection, p. 9).

Within the triadic conception of representation, O’Brien identifies two hurdles for a resemblance theory that still need to be overcome: it must be shown how the theory can be compatible with physicalism, and it must be secured that the theory does not leave content indeterminate (ibid., pp. 9–11).

In order to secure the compatibility of a resemblance theory of content determination with physicalism, O’Brien turns away from the notion of *first-order resemblance* and instead makes use of *structural* or *second-order resemblance* (ibid., pp. 10–11). He thus avoids the seemingly naive and implausible thesis that

² It might be said that this understanding of representation is plausible for paradigmatic cases of representation, like the representation of material objects, but that it is less clear whether it is also fit to capture cases that differ strongly from those paradigms, like the representation of abstract “objects”.

A similar worry, which has been pointed out to me by an anonymous reviewer, concerns cases of *fictional* representations, e.g., future events. For this specific example, she/he suggests that we allow for the causal chain to be reversed. This would make the representandum part of the final event. However, this solution applies only to those cases of fictional representation where the representandum does *not yet* exist, but leaves the majority of cases of fictional representation inexplicable.

³ I will, inspired by O’Brien’s terminology, keep referring to the representandum as the represented *object*. The term “object” is thereby used in a very wide sense and not intended to be restricted to single material objects. What exactly can take the place of an object in O’Brien’s story of representation is yet to be determined.

⁴ This question presupposes that the established representation is actually correct and not a case of misrepresentation.

mental representations must actually share properties with what they represent (*ibid.*, p. 10). Resemblance is taken to a more abstract level where, for example, something red can be mentally represented *with the representation resembling the representandum*, but *without* them both sharing the property of being red (*ibid.*, pp. 10–11).

The second hurdle might seem redundant at first glance. Explaining how content is determined is basically the job description of a theory of content determination. It is still worth mentioning this as an obstacle, however, because the reliance on second-order resemblance makes this job look particularly difficult: second-order resemblance is too easily established. If a set of mental representations second-order resembles a pattern of colour shades, it might in virtue of the same relational organization also second-order resemble a pattern of locations in a two-dimensional space. Nevertheless, O'Brien trusts that within the triadic conception of representation, second-order resemblance will do the job. The idea is that some of the possibilities for the content of a vehicle that are left open by second-order resemblance are ruled out in the process of interpretation—interpretation is “content-limiting” by “anchoring vehicles” in “domains” (O'Brien *this collection*, p. 11). The preexisting behavioural dispositions influence the newly developing ones, so that they are not directed towards all domains with a specific relational organisation, but towards a selection of these.

So far, O'Brien has provided us with an interesting account of how mental phenomena are causally efficacious in virtue of their representational contents: the property

x has representational content

is analyzed as the property

x results from a causal process that brings about behavioural dispositions towards the object that triggered the causal process.

These behavioural dispositions, given their respective stimuli, can now yield causal effects.

But if we took this as O'Brien's only accomplishment, we would miss the most interesting part of his argument. Furthermore, we would take the second step before the first.

3 Content and causation: from a question to a problem

So far, I have interpreted O'Brien's formulation of the content causation problem (“How can mental phenomena be causally efficacious of behaviour in virtue of their representational contents if these contents are not determined by intrinsic properties of the brain?” O'Brien *this collection*, p. 2) as something along the lines of “How does mental causation in virtue of representational content work, if not in the way we already know it sometimes works?” In so doing, one already engages in an interesting discussion about content causation. But closer examination reveals that this understanding of the problem is an oversimplification. The content causation problem is supposed to be much more severe. It is not a problem about finding alternative explanations to the ones we already have, but about the *consistency* of all available explanations. A better understanding of the problem will help us to evaluate whether O'Brien's suggestions, which are doubtlessly interesting, are really motivated by the problem at hand.

The three theses, that (i) the causal efficacy of mental phenomena is grounded in their representational contents, that (ii) these “are not determined by the intrinsic properties of the brain” (O'Brien *this collection*, p. 2), and that (iii) the brain's causal efficacy for behaviour is grounded only in its *intrinsic* properties, supposedly “form an inconsistent triad” (*ibid.*, p. 2). Yet, strictly speaking, they are not inconsistent: why not say that what we do in theses (i) to (iii) is gathering information about (human) behaviour—our object of enquiry—but on two levels of description? On both levels, we attempt, metaphorically speaking, to travel back along the causal chain that brings behaviour about. On the one level, we then discover that intrinsic properties of the brain are responsible for it to cause behaviour (*ibid.*, p. 2, thesis 3). On another level, we trace behaviour back to

mental phenomena, which owe their capacity to cause it to their representational contents (*ibid.*, p. 2, thesis 1). Why not assume that these two levels both provide us with (true) formulations of what is happening, but which—since they depend on two conceptual frameworks that are not intertranslatable—must be regarded as *nomologically incommensurable*? If so, they could never both be part of a unified causal theory (see Davidson 1970). This picture seems perfectly intelligible at first. But on both levels, we talk about causes of (presumably the same) behaviour.

Within a physicalist framework, we take behaviour to fall, in the end, under the description of a physical event. As such, it is subject to *the principle of causal closure*: if it is caused, it has a sufficient physical cause (Kim 1989, p. 43). Hence, we should pay close attention to the fact that “our best neuroscience informs us that the changes to musculature that constitute our behavioural responses are wholly determined by the intrinsic properties of the brain to which they are causally connected” (O’Brien *this collection*, p. 2). Thus it is not only assumed that the brain *can be* causally efficacious of behaviour in virtue of its intrinsic properties, but also that *whenever* behaviour is caused, it is *always* caused in the brain and in virtue of the brain’s intrinsic properties. Yet mental phenomena, which are of a non-physical kind, are also mentioned in (i) as a cause of behaviour. But with brain states already providing a sufficient cause for behaviour, what role in causation can they possibly play (Kim 1989, pp. 43–44)? As long as we cannot answer question, we are forced to reject the possibility that behaviour is *over-determined*, or, in other words, we are forced to accept both mental phenomena and brain states as two *separate* causes of behaviour. So causally efficacious mental phenomena should be reducible to the physical cause already provided by the states of the brain, or we must conclude that they are not a cause at all. In the latter case, this would mean that we would have to deny “that our beliefs and desires, and our perceptions and thoughts ultimately have a causal impact on our behaviour” (O’Brien *this collection*, p. 1) and we would

have to accuse every discipline accepting this tenet—O’Brien names folk psychology and the computational theory of mind (*ibid.*, p. 2)—of operating with a faulty ontology, pointing out causes that do not really exist.⁵

Now we have made explicit a metaphysical constraint that was left implicit in the formulation of the content causation problem: mental properties and all their capacities, e.g., their capacity to cause behaviour, must be reducible to properties of the physical brain. Knowing this, we see where the supposed inconsistency comes from: we traditionally think of representational content not as determined by the brain’s intrinsic properties, but rather as determined by what it is about (O’Brien *this collection*, p. 2–3). But if the content causation problem dares us to integrate these two things, namely the description of mental phenomena as causing behaviour in virtue of their representational contents and our theory of the same behaviour being caused by processes in the brain in virtue of the brain’s *intrinsic* properties, then we might conclude with O’Brien:

A solution to the content causation problem requires something that *prima facie* appears impossible: an explanation of the relational character of mental content that invokes only the intrinsic properties of the brain. (*this collection*, p. 3)

This is what turns the content causation problem as formulated by O’Brien from an interesting question into an urgent philosophical problem. The apparent impossibility of a solution relies on the idea of a sharp distinction between relational and intrinsic properties. If our best shot at understanding whatever we describe as the “relational character” of representational content is to understand it as a relational property,⁶ then its irreducibility to intrinsic properties of the brain is built into it—and so is the insolvability of the content causation problem, given the metaphysical constraint just men-

⁵ The threat lurking in the background is, of course, eliminative materialism (cf. Churchland 1981).

⁶ Remember that this kind of property is even referred to as “not determined by the intrinsic properties of the brain” (O’Brien *this collection*, p. 2).

tioned. Nevertheless, O'Brien aims to provide a solution.

4 The relational character of representational content

We see now that an attempt to solve the content causation problem must address the question of how the specific character of representational content can be analyzed in a way that invokes only the intrinsic properties of the brain (instead of being understood as “being a relational property and not an intrinsic property of the brain”). However, O'Brien advertises a theory of content determination that draws on second-order similarity between mental vehicles and the outside world as necessary for a solution to the content causation problem. In fact, he admits that “all theories of mental representation, in their efforts to explain the relational character of mental content, are forced to invoke world-mind relations of some kind”, where the latter term seemingly refers back to “relations that incorporate factors beyond the brain” (O'Brien [this collection](#), p. 12). But how does this relate to the explicit goal of providing “an explanation of the *relational* character of mental content that invokes only the *intrinsic* properties of the brain”, which would only “*prima facie* [appear]” to be, but not—as the content causation problem is supposed to have a solution—actually *be* “impossible” (O'Brien [this collection](#), p. 3)?

The answer to this might lie in a view which can be found in O'Brien & Opie:

Von Eckardt observes that the triadicity of representation in general, and mental representation in particular, can be analysed into two dyadic component relations: one between representing vehicle and represented object (which she calls the *content grounding* relation); the other between vehicle and interpretation [...] This suggests that any theory of mental representation must be made up of (at least) two parts⁷: one that explains how the content

of mental vehicles is grounded, and a second that explains how they are interpreted. (2004, p. 5)

When O'Brien writes that every theory of representational content, including his own, must make use of factors extrinsic to the brain, he most likely refers to the content grounding relation—in his case, second-order resemblance. Yet when he promises us “an explanation of the relational character of mental content that invokes only the intrinsic properties of the brain” (O'Brien [this collection](#), p. 3), the scientific explanation mentioned most likely refers to the other part of the theory of mental representation: the internalist theory of interpretation. If O'Brien takes it that only this theory, which provides us with a reconstruction of the causal processes involved in mental representation, needs to be presented in terms of intrinsic properties of the brain, then he provides an account within the “narrow content program” (*ibid.*, p. 3): this research program accepts the thesis that “[t]he representational contents of mental phenomena are not determined by the intrinsic properties of the brain” (*ibid.*, p. 2) but—quite plausibly, I think—relaxes the metaphysical constraint made explicit in section 3 insofar as it only demands “an account of mental phenomena according to which (*at least the causally relevant component of*) *their representational properties* are determined by intrinsic properties of the brain” (*ibid.*, p. 3, my emphasis).

If this is a correct reconstruction of O'Brien's steps towards a solution to the content causation problem, then he has reached his goal if he:

- a) has provided an account of the causally-relevant components of representation that makes use of only the intrinsic properties of the brain, and
- b) can make sure that this account still deserves to be called an account of representation, i.e., captures the specific characteristics of representational content that we have so far called “relational”.

O'Brien claims that “[t]he insight offered by triadicity is that the relational character of mental

⁷ A third relation that one might want to look at when “taking apart” the triadic relation of representation is the relation between interpretation and represented object.

content is to be discharged ultimately in terms of our behavioural dispositions towards features of the world” ([this collection](#), p. 12). While it might not be clear at first sight why this provides a solution to the problem at hand, I am convinced that a view of dispositions as second-order properties—such as the property of having a property that becomes relevant for the causing of a certain manifestation once a certain stimulus is provided—helps us to see how O’Brien’s account provides a solution. I hope to have shown in section 2 how the adoption of the view that dispositions are second-order properties fits into his picture of representation. Thus I believe that it allows us to regard the first of the two requirements mentioned above as fulfilled: I see no reason why such a second-order property should not be understood as an intrinsic property of the brain. Furthermore, I hold that this view allows us to regard the second requirement as fulfilled, too: the dispositions in question seem to deserve the label “relational” insofar as, when combined with a certain stimulus, they are manifested in terms of overt, observable behaviour of a biological organism. When so manifested, they turn into concrete chains of events linked by causal *relations*. One can now argue that these potential relations are what let us intuitively characterize representation as relational. So understood, O’Brien’s explanation does justice to the project of providing an account that captures the specific character that makes representational content deserve the label “representational”, but without characterizing its causally efficacious components as being relational properties.

5 The content causation problem and second-order resemblance relations

If this is to be seen as a successful analysis of representational content in terms of intrinsic properties of the brain, we can conclude that the triadic picture alone—with its analysis of the relational character of representational content in terms of dispositions—already solves the content causation problem. Hence, this problem, by itself, offers criteria that could be turned into an argument for or

against any specific theory of content determination.

Such a theory, as I understand it, serves two purposes: it explains how the contentful vehicles of which the theory of interpretation makes use are individuated, and it explains “how relations between mental vehicles and their represented objects can endow subjects with the capacity to respond selectively to those very features of the world” (O’Brien [this collection](#), p. 6). It thus provides the background information necessary for understanding why the theory of interpretation is able to do what is required of it. The second desideratum is only made clear if the triadic account of content causation is adopted. According to O’Brien, it can only be fulfilled if we adopt the resemblance theory of content determination, because “[t]he obtaining of causal relations between external conditions and inner vehicles cannot explain how the latter endow systems with the capacity to respond in a discriminating fashion towards the former” (*ibid.*, p. 12), whereas “resemblance does offer some prospect of a solution to the content causation problem. The key here is that the mere obtaining of the resemblance ensures that the former have properties that can be exploited to shape the behavioural dispositions of cognitive systems towards the latter” (*ibid.*, p. 12).

Let us recapitulate the steps that took us from the content causation problem to the second-order resemblance theory of content determination. The content causation problem motivates the triadic account of representation if we assume that it is a problem about reduction and if we assume that there is a sharp distinction between relational and intrinsic properties that forbids an analysis of the former in terms of the latter, thus preventing the reduction of a theory of causally efficacious mental phenomena to a theory of brain-based causation of behaviour. The triadic account of representation solves this problem by showing us that we need not assume that representational content owes its specific character to relational properties. Dispositions, understood as non-relational second-order properties, do justice to our concept of “relational character”. The triadic ac-

count then needs to be enriched by a theory of content determination, and its use of the concept of a “disposition” leads to a new requirement: to explain how inner vehicles can enable a subject to respond selectively towards external objects. Supposedly, only second-order resemblance can fulfill this requirement. Thus understood, a second-order resemblance theory of content-determination is only indirectly relevant to a solution of the content causation problem.

Yet I would like to point out a way in which the second-order resemblance theory itself relates to the content causation problem. Second-order resemblance between vehicles and objects tells us that there is a mapping from objects to vehicles that is pattern-preserving or, in other words, some objects and some vehicles are alike in some of their relational properties. Nevertheless, the kind of pattern involved is to be “sustained by constraints inherent in the vehicles, rather than being imposed extrinsically” (O’Brien [this collection](#), p. 11, footnote 11). The relations constituting a structure or pattern collectively supervene on the distribution of intrinsic properties of objects and vehicles, although individual extrinsic (and specifically relational) properties of objects and vehicles do not. This strategy of explanation is in principle also available to our understanding of content: contents, fixed by a structural organisation of vehicles, are relational in the same, completely unproblematic sense. Representational contents may not *individually* be determined by intrinsic properties of the brain, but there is a sense in which they are so collectively. But this might count as evidence against the second thesis of the content causation problem: that “[t]he representational contents of mental phenomena are not determined by the intrinsic properties of the brain” (*ibid.*, p. 2). One might then even say that content is *not relational at all*, and that the puzzle that actually troubles us is the question of how representations can have something to which they are applied, namely a “*target*” (Cummins 1996, p. 8).⁸ This could still be accounted for by the triadic picture of represent-

ation, but it would then not amount to *solving* the content causation problem, but to rejecting it.

6 Conclusion

In his target article, Gerard O’Brien addresses the question of “how the specifically representational properties of mental phenomena can be causally efficacious of behavior” ([this collection](#), p. 12). When he does so, there are two parts of the problem to be considered: the first is explaining how mind matters, and the second is showing how an answer can prevail in the light of the content causation problem. Considering the first part in isolation, O’Brien provides an interesting answer. He translates our talk of representation into causal vocabulary, thereby making it possible to reach a concept of causally efficacious representational content. In order to understand how O’Brien’s account needs to be assessed with regard to the second part, one first needs to reconstruct which background assumptions make the content causation problem so pressing.

I am convinced that the issues of *reduction* and the *relational/intrinsic property distinction* need to be addressed in order to understand whether and how the content causation problem can motivate an account like O’Brien’s. His account takes as a starting point that representational content has a relational character, but should not be understood in terms of relational properties. Rather, as we have seen, it should be understood in terms of dispositions—which can, if manifested, establish causal relations, but are not relational by themselves. I hope to have provided a reconstruction of how this starting point is used to reach the conclusion that, as O’Brien formulates it, “resemblance theories appear obligatory, since they alone offer some prospect for explaining how mind matters” ([this collection](#), p. 12).

If this is correct, there remains one question: whether resemblance theories of the proposed kind might themselves indicate that the content causation problem rests on a mistake. The problem presupposes further problems about the role of relational and intrinsic proper-

⁸ I owe this point to an anonymous reviewer.

ties that need not be addressed in order to account for the causal efficacy of representational content. The content causation problem's not arising in the first place would, of course, not undermine O'Brien's highly interesting account. It is only that this problem could no longer be used to motivate the argumentative steps he takes. Still, his account is illuminating for many other reasons, such as translating mysterious talk about "being about" into naturalistic terminology. However, whether we can regard the content causation problem as solved or rather as successfully rejected is not clear; but instead of worrying about this problem, we might now turn towards the details of O'Brien's account. An interesting starting point for such further inquiry might be to try to reach a better understanding of the role and the kind of *dispositions* and *vehicles* involved in causal processes, for they form two of the key concepts in O'Brien's theory.

References

- Choi, S. & Fara, M. (2014). Dispositions. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy (Spring 2014 Edition)*. <http://plato.stanford.edu/archives>.
- Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *The Journal of Philosophy*, 78 (2), 67-90.
- Cummins, R. (1996). *Representations, targets, and attitudes*. Cambridge, MA: MIT Press.
- Davidson, D. (1970). Mental events. In L. Foster & J. W. Swanson (Eds.) *Experience and theory* (pp. 79-101). Atlantic Highlands, NJ: Humanities Press.
- Kim, J. (1989). The myth of nonreductive materialism. *Proceedings and Addresses of the American Philosophical Association*, 63 (3), 31-47.
- O'Brien, G. (2015). How does mind matter? - Solving the content causation problem. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- O'Brien, G. & Opie, J. (2004). Notes towards a structuralist theory of mental representation. In H. Clapin, P. Staines & P. Slezak (Eds.) *Representation in mind: New approaches to mental representation* (pp. 1-20). Amsterdam, NL: Elsevier.

Rehabilitating Resemblance Redux

A Reply to Anne-Kathrin Koch

Gerard O'Brien

Anne-Kathrin Koch's insightful commentary places a great deal of pressure on the connection between my deployment of the triadic analysis of representation to solve the content causation problem and my contention that it makes mandatory the rehabilitation of the resemblance theory of mental content determination. She argues that if the relational character of mental content can be captured in terms of brain-based behavioural dispositions, as I claim, then this manoeuvre in its own right solves the content causation problem and hence offers no support for resemblance or any other theory of content determination. In this reply, I argue that the relation between the proposed solution to the content causation problem and the resemblance theory of content determination is stronger than Koch allows.

Keywords

Content determination | Mental causation | Mental content | Mental representation | Resemblance

Author

Gerard O'Brien

gerard.obrien@adelaide.edu.au
University of Adelaide
Adelaide, Australia

Commentator

Anne-Kathrin Koch

anne-kathrin.koch@gmx.de
Johannes Gutenberg-Universität
Mainz, Germany

Editors

Thomas Metzinger

metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Introduction

There is a paradoxical air surrounding mental content. On the one hand we take it to be a localized property of our minds—of our mental states—distinct from the world in which we are embedded. Yet on the other hand, it is the means by which our minds reach out and make “cognitive contact” (Kriegel 2003) with this surrounding environment. How is such action-at-a-distance possible? The standard solution to this conundrum is to assume that the **relational character** of mental content can be explained by the fact that mental content is a **relational property** of our mental states. This line of thought leads to **content externalism**, accord-

ing to which mental content is determined in part by factors beyond our heads. But once content externalism is combined with a couple of unexceptional theses about (i) the role of content in mental causation and (ii) the brain-basis of the causal determinants of behaviour, we encounter the **content causation problem**—the problem of explaining how the content of mental states can be causally efficacious of behaviour when it doesn't supervene on what's in our heads.

The solution I offered in my target paper was to sever the connection between the relational character of mental content and the as-

sumption that the latter is a relational property of our mental states (O'Brien [this collection](#)). My suggestion was that unlike a **dyadic** story that seeks to explain representation solely in terms of relations between vehicles and their represented objects, a **triadic** account of representation opens up space to explain the relational character of mental content in terms of brain-based behavioural dispositions—specifically, dispositions to respond selectively to specific features of the external environment. According to this triadic account, the **aboutness** of mental content is not some mysterious relational property that brings our minds into contact with various aspects of the surrounding environment; it is the relatively straightforward cognitive capacity, bestowed by the intrinsic properties of our brains, to regulate our behaviour in response to specific environmental conditions.

In her insightful commentary, Anne-Kathrin Koch, after carefully rendering explicit some of the background assumptions on which I rely, focuses on the connection between the proposed solution to the content causation problem and my further contention that it makes mandatory the rehabilitation of the resemblance theory of mental content determination (Koch [this collection](#)). Her counter claim is that if the relational character of mental content can be successfully captured in terms of the brain's behavioural dispositions, then this manoeuvre in its own right solves the content causation problem and hence offers no support for resemblance or any other theory of content determination. In this reply, I will show that the relation between the proposed solution to the content causation problem and the resemblance theory of content determination is stronger than Koch allows.

2 Rejecting resemblance (and content causation)

The great insight of Charles Sanders Peirce's analysis of representation is his claim that aboutness can't be explained solely in terms of relations between representing vehicles and represented objects (Hardwick 1977). Instead, vehicles are about their objects in virtue of hav-

ing a certain kind of effect on a cognitive subject—specifically, vehicles either trigger thoughts about objects (in cases of public representation) or they engender behavioural dispositions towards them (in cases of mental representation). According to Peirce, it is this additional relatum—known as **interpretation**—that renders representation triadic.

But once interpretation is added into the representational mix, it has the potential to overwhelm any content-grounding relations that may obtain between vehicles and objects. This is the thread that Koch astutely pulls on in her commentary. To the extent that one can appeal to the manner in which a representing vehicle modifies a subject's behavioural dispositions in order to capture the relational character of content, it seems as though one can also appeal to these dispositions to fix the content of this vehicle. In short, the triadic account would appear to make those theories of content determination that appeal to vehicle-object relations—such as resemblance—redundant (or, at least, “only indirectly relevant”, as Koch charitably puts it; [this collection](#), p. 8).

Koch is not alone in drawing out this consequence from the triadic nature of representation. It is precisely this idea about the role of behavioural dispositions in content fixation that forms the foundation of the instrumentalist approach to mental representation that Daniel Dennett has defended over many years (1978; 1987). Dennett was one of the early proponents of triadicity, insofar as he argued that it was only in virtue of their roles in cognitive systems that representing vehicles can be interpreted as bearers of information:

There is a strong by tacit undercurrent of conviction [...] to the effect that only by being rendered explicit [...] can an item of information play a role. The idea, apparently, is that in order to have an effect, in order to throw its weight around, as it were, an item of information must weigh something, must have a physical embodiment [...]. I suspect, on the contrary, that this is almost backwards. [Representing vehicles]... are by themselves quite inert as

information bearers [...]. They become information-bearers only when given roles in larger systems. (Dennett 1982, p. 217)

Dennett has also famously argued that the consequence of taking the triadic account seriously is the rejection of any story that takes mental content to be determined independently of a cognitive creature's patterns of behaviour.

Dennett's instrumentalist approach to mental representation, however, has another famous consequence. If the full burden of content determination falls on the shoulders of interpretation—if, that is, it is a cognitive system's behavioural dispositions ultimately fix the content of its representing vehicles—then content is a product of cognition, not an ingredient, and hence cannot be casually implicated in the production of behaviour.

This last point, of course, represents Dennett's own solution to the content causation problem: he abandons the thesis that mental phenomena are causally efficacious of behaviour in virtue of their representational contents (see O'Brien [this collection](#), fn. 1). This is also what Koch is hinting at when she indicates that my proposal to invoke the triadic account of representation might be better interpreted as **rejecting** the content causation problem rather than **solving** it (Koch [this collection](#), p. 8). That is, far from showing that rehabilitation of the resemblance theory of content determination is mandatory, her (implicit) objection is that my proposed solution to the content causation really shows that there is no such thing as content causation in the first place.

3 Rehabilitating content causation (and resemblance)

To reiterate, the problem associated with the triadic analysis of representation is that once behavioural dispositions are invoked in order to explain the relational character of mental content, they threaten to overwhelm any other story about mental content determination. But if mental content is determined by such behavioural dispositions, it can't play a robust causal role in their production. In short, the triadic ac-

count seems to suggest that there is no content causation **problem** because there is no **content causation**.

In this context, however, it is pertinent to note that, despite his insistence that representation is triadic, Peirce expends a good of effort investigating the relations between representing vehicles and represented objects. His analysis of public forms of representation famously yields three different kinds of vehicle-object relations—convention, causation, and resemblance—associated with symbols, indexes, and icons, respectively (see Hardwick 1977 and Von Eckardt 1993, Ch. 4). If content determination is ultimately just a matter of interpretation, why would Peirce have been so bothered about these vehicle-object relations?

The answer, of course, is that Peirce was concerned not just with the fact that public representing vehicles effect interpretations in cognitive subjects, but with **how** they do so. The point here is that interpretation isn't magic—it requires explanation. Consider, for example, Leonardo da Vinci's **Mona Lisa**. According to the triadic story, the painting that hangs in the Louvre is not about anything on its own. Its standing as a representing vehicle hinges on its capacity to trigger interpretations in cognitive subjects. When we look at this painting it causes us to think about a dark-haired woman with a famously enigmatic smile. But what is it about this painting that endows it with this capacity? Part of the explanation here invokes our recognition of the resemblances between the painting and a woman with a certain kind of physical appearance. The painting wouldn't have the same impact on us if these resemblances didn't obtain. So a complete account of the painting's aboutness must go beyond the fact that it triggers certain thoughts in us to include an explanation of how it does so. And it is here that vehicle-object relations such as resemblance are compulsory.

The general lesson to take away from this (far too brief) analysis is that the interpretation of public forms of representation cannot be disconnected from the cognitive subject's (conscious or unconscious) recognition of what are generally known as **content grounding** rela-

tions between vehicles and represented objects. And what goes for the interpretation of public representing vehicles also goes for the interpretation of mental vehicles. On the triadic story being entertained here, mental vehicles, just like the **Mona Lisa**, aren't about anything considered in isolation. Their aboutness is a consequence of the multifarious behavioural dispositions they create in us towards selective features of the world—dispositions to physically interact with these features, for example, or to make utterances about them. Since this form of interpretation likewise isn't magic, a complete account of mental representation must explain how mental vehicles establish these behavioural capacities. And just as with the case of public representing vehicles, it is impossible to do this without recourse to content-grounding relations (something that is demonstrated by even exceedingly simple representation-using devices such as the humble thermostat—see O'Brien [this collection](#), pp. 7–9).

Precisely because content-grounding relations must be invoked to explain how the former endow cognitive systems with behavioural dispositions towards the latter, content causation is back in business. But what kind of vehicle-object relations can turn this trick? This, of course, was one of the central questions that animated much of my discussion in the target paper (O'Brien [this collection](#)). Of the three grounding relations that Peirce found to be implicated in public forms of representation—convention, causation, and resemblance—the first is widely assumed to be unavailable for mental representation since it violates the naturalism constraint.¹ Despite its popularity in contemporary philosophy, the second, I argued at some length, is actually powerless to explain how mental vehicles create the requisite behavioural

dispositions ([this collection](#), pp. 6–7). This just leaves us with resemblance. Fortunately, this third vehicle-object relation is up to the task, or at least so my argument went, since the structural properties of mental vehicles that ground second-order resemblance relations can be exploited to shape the behavioural dispositions of a cognitive system towards worldly objects ([this collection](#), p. 8). This is where the rubber of resemblance meets the road of content causation. And it is why a resemblance theory of content determination is mandatory if we are to explain why mind matters.

References

- Dennett, D. (1978). *Brainstorms*. Cambridge, MA: MIT Press.
- (1982). Styles of mental representation. *Proceedings of the Aristotelian Society, New Series*, 83, 213–226.
- (1987). *The intentional stance*. Cambridge, MA: MIT Press.
- Hardwick, C. (Ed.) (1977). *Semiotics and signification: The correspondence between Charles S. Peirce and Victoria Lady Welby*. Bloomington, IN: Indiana University Press.
- Koch, A.-K. (2015). Does resemblance really matter? – A commentary on Gerard O'Brien. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Kriegel, U. (2003). Real narrow content. *Mind and Language*, 23 (3), 304–328.
[10.1111/j.1468-0017.2008.00345.x](https://doi.org/10.1111/j.1468-0017.2008.00345.x)
- O'Brien, G. (2015). How does mind matter? - Solving the content causation problem. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Von Eckardt, B. (1993). *What is cognitive science?* Cambridge, MA: MIT Press.

¹ This is the requirement that mental representation be explained without appeal to further forms of representation. If a vehicle is related to its object by convention, the cognitive subject must deploy a *rule* that specifies how the vehicle is to be interpreted. In the case of non-mental representation, where for example the vehicle is a word in a natural language, the application of such a rule is a cognitive achievement that must be explained in terms of processes defined over mental representing vehicles. When this same account is applied to mental vehicles, therefore, it would seem to generate an infinite regress of further representing vehicles, and hence interpretation is never achieved (see Von Eckardt 1993, p. 206).

Conscious Intentions

The Social Creation Myth

Elisabeth Pacherie

What are intentions for? Do they have a primary purpose or function? If so, what is this function? I start with a discussion of three existing approaches to these questions. One account, associated with Michael Bratman's planning theory of agency, emphasizes the pragmatic functions of intentions: having the capacity to form intentions allows us to place our actions more firmly under the control of deliberation and to coordinate our actions over time. A second account, inspired by Elizabeth Anscombe's theory of intentions, emphasizes their epistemic function and their contribution to self-knowledge. A third account, developed by David Velleman, suggests instead that the capacity for intentions may be an accident or a spandrel, that is, a byproduct of some more general and fundamental endowments of human nature. I argue that these accounts are at best partial and largely overlook two important dimensions of intention. I introduce and motivate a further pragmatic function of intentions, namely their role in the control and monitoring of ongoing action and argue that acknowledging the existence and importance of this function allows us to plug gaps in these accounts. I further argue that this pragmatic function of intentions plays a crucial role in contexts of joint action where agents must align their representations in order to coordinate their actions towards a joint goal. I speculate that a capacity for conscious control might have become established because of the role it served in solving inter-agent coordination problems in social contexts and because of the benefit conferred by the forms of cooperation it thus made possible.

Keywords

Action coordination | Conscious action control | Intention | Joint action | Planning | Representational alignment | Self-knowledge

1 Introduction

What are conscious intentions for? What do we gain from having a capacity for intentions as opposed to simply a capacity for desire-belief motivation? Do intentions have a function not just in the sense that they have a causal role but in the normative sense in which having this function confers benefits on intention-forming creatures that explain why these creatures have this capacity. In other words, do intentions have a teleofunction? Is there something they are for? And if so, what is this teleofunction?

Roughly, the notion of intention is that of a mental state that represents a goal (and means to that goal) and contributes through the guidance and control of behavior to the realization of what it represents. Thus, my intending to go to my office will control and guide my behavior (e.g., leaving my house, taking the bus, walking from the bus stop to my office), thus contributing to the realization of the goal represented by the intention. Many philosophers hold the view that if we do something intentionally, we must be aware of what we are doing.

Author

Elisabeth Pacherie

elisabeth.pacherie@ens.fr

Ecole Normale Supérieure
Paris, France

Commentator

Andrea R. Dreßing

andrea.dressing@uniklinik-freiburg.de

Klinik für Neurologie und
Neurophysiologie
Universitätsklinikum Freiburg
Freiburg, Germany

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

Therefore, they consider that it is of the essence of intentions to be conscious. I have argued elsewhere (Pacherie 2008) in favor of a notion of motor intentions whose contents may not always be accessible to consciousness. On my view then, the phrase “conscious intentions” need not be pleonastic. Here, however, my focus will be on intentions qua conscious states and I will use “conscious intentions” and “intentions” interchangeably.

In his 2007 paper, “What good is a will?”, David Velleman considers the question whether the human will, understood as the capacity for (conscious) intentions, has a purpose or teleofunction. He discusses two accounts that assume that the will has a purpose but disagree on what this purpose is. On one account, associated with Bratman’s planning theory of agency, the primary function of intentions is pragmatic: having the capacity to form prior intentions is good because it allows us to place our actions more firmly under the control of deliberation and to coordinate our actions over time. On the other account, inspired by Anscombe’s theory of intentions, the primary function of intentions is epistemic. Intentions are good because they provide self-knowledge: an intention on which one acts provides us with a special kind of knowledge of what one is doing.

David Velleman is himself skeptical that the attitude of intention has a teleofunction. Rather, he suspects that the human will is an accident or a spandrel, that is a byproduct of some more general and fundamental endowments of human nature. Velleman suggests, however, that our hypotheses about the origins of the will, including his own, must be closer to creation myths than to scientific theories. Talk of myths, of course, has both negative and positive connotations. On the negative side, myths are, if not downright false or unfounded, at least ultimately unverifiable. On the positive side, myths are dramatization devices that serve to highlight, and make sense of, the value or function of a practice, of an institution or, in the case at hand, of a cognitive capacity. Here, I will offer my own creation myth for intentions, a myth that emphasizes the social dimension and social function of conscious intentions. The

main claim I will defend is that having conscious intentions is a good thing in large part because it facilitates coordination and cooperation with others and because cooperation is itself fitness enhancing. My aim in proposing this social creation myth is not to entirely displace other creation myths, but rather to complement them, to highlight an important facet of conscious intentions that traditional philosophy of action has tended to neglect and to plug some holes in the stories told in other myths.

Here’s how I will proceed. In section 2, I will present the two creation myths considered and rejected by Velleman and discuss some difficulties they raise. In section 3, I will discuss Velleman’s own creation myth. In section 4, I will introduce and motivate a pragmatic function of intention largely overlooked by these creation myths, namely their role in the control and monitoring of ongoing actions. In section 5, I will tell my own social creation myth. I’ll argue that this pragmatic function of intentions plays a crucial role in contexts of joint action where agents have to align their representations in order to coordinate their actions towards a joint goal. I’ll speculate that the main evolutionary benefit conferred by a capacity for conscious intentions is that it enables a considerable increase in the possibilities for joint action and cooperation.

2 Two teleological creation myths

Velleman (2007) points out a methodological assumption common in functionalist psychology, namely the assumption that our attitudes or cognitive faculties have a function not just in the sense that they have a causal role but in the sense that they have a purpose, something they are designed to do and thus ought to do. Functions in this latter sense are commonly called teleofunctions. This methodological assumption needs not entail a belief in some intelligent designer. Instead, it can be cashed out by appealing to evolutionary theory and to natural selection as a blind designer. Typically, the evolutionary story goes like this: a trait has the teleofunction of producing effect E just in case producing this effect conferred some benefit that

contributed to the reproductive success of organisms endowed with the trait and, thereby, to the propagation of the trait itself. This methodological assumption, when it guides our inquiry into intentions, leads us to take the question what intentions are for, i.e., what purpose are they meant to serve, as necessarily meaningful and demanding an answer.

Velleman discusses two teleological stories meant to answer this question. He links the first story to [Bratman's](#) theory of intentions (1987) and the second to [Anscombe's](#) theory (1963). I start with the story inspired by Bratman's theory.

We are, in Bratman's words, planning agents regularly making more or less complex plans for the future and guiding our later conduct by these plans. This planning ability appears to be if not unique to humans at least uniquely developed in the human species. People can, and frequently do, form intentions concerning actions not just in the near but also in the distant future. Why should we bother forming future-directed intentions? What purposes can it serve? What benefits does it bring us? What features of future-directed intentions allow them to serve these purposes?

Bratman offers two complementary answers to that challenge. The first stems from the fact that we are epistemically limited creatures: our cognitive resources for use in attending to problems, gathering information, deliberating about options and determining likely consequences are limited and these processes are time consuming. As a result, if our actions were influenced by deliberation only at the time of action, this influence would be minimal as time pressure isn't conducive to careful deliberation. Forming future-directed intentions makes advance planning possible, freeing us from that time pressure and allowing us to deploy the cognitive resources needed for successful deliberation. Second, intentions once formed commit us to future courses of action, thus making the future more predictable and making it possible for agents to coordinate their activities over time and to coordinate them with the activities of other agents. Making deliberation and coordina-

tion possible are thus the two main benefits that accrue from a capacity to form future-directed intentions.

What makes it possible for future-directed intentions to yield these benefits is, according to Bratman, the fact that they essentially involve commitments to action. Bratman distinguishes two dimensions of commitments: a volitional dimension and a reasoning-centered dimension. The volitional dimension concerns the relation of intention to action and can be characterized by saying that intentions are "conduct-controlling pro-attitudes" ([Bratman 1987](#), p. 16). In other words, unless something unexpected happens that forces me to revise my intention, my intention today to go shopping tomorrow will control my conduct tomorrow. The reasoning-centered dimension of commitment is a commitment to norms of practical rationality and is most directly linked to planning. What is at stake here are the roles played by intentions in the period between their initial formation and their eventual execution. First, intentions have what Bratman calls a characteristic stability or inertia: once we have formed an intention to A, we will not normally continue to deliberate whether to A or not. In the absence of relevant new information, the intention is rationally required to resist reconsideration: we will see the matter as settled and continue to so intend until the time of action. Intentions are thus terminators of practical reasoning about ends or goals. Second, during this period between the formation of an intention and action, we will frequently reason from such an intention to further intentions. For instance, we reason from intended ends to intended means or to preliminary steps. When we first form an intention, our plans are typically only partial, but if they are to eventuate into action, they will need to be filled in. Thus intentions are also prompters of practical reasoning about means. Third, because intentions are commitments to action, our intentions should be jointly executable. Finally, taken together the volitional and the reasoning-centered dimensions of commitments help explain how intentions can promote coordination. They provide support for the expectation that agents will act as they intend to and these ex-

pectations are central in turn to both inter- and intra-personal coordination. In particular, this is what motivates the rational agglomerativity requirement on intentions, i.e., the requirement that my intentions be jointly executable.

The benefits that accrue from a capacity for intentions are, ultimately, pragmatic benefits. As Bratman puts it, future-directed intentions “enable us to avoid being merely time-slice-agents” (1987, p. 35). Instead of constantly starting from scratch in our deliberations and simply weighing current belief-desire reasons, intentions allow us to become temporally extended agents. They provide a background framework that allows us to expand the temporal horizon of our deliberation while at the same time narrowing its scope to a limited set of options. In so doing they contribute in the long run to our securing greater desire-satisfaction than simple desire-belief practical reasoning would.

Velleman (2007) sees three main problems with Bratman’s pragmatic account of what intentions are for. The first problem concerns the status and role of present-directed intentions. On Bratman’s account, a future-directed intention requires a present-directed intention to convey its motivational force and guide the action once the time to act is seen to have arrived. Bratman identifies no further role or function of present-directed intentions beyond conveying the motivational potential of future-directed intentions. At the same time, he insists that intentional actions, whether or not they are preceded by future-directed intentions, always involve present-directed intentions. This leaves us with a potentially large class of spontaneous intentional actions that involve present-directed intentions but are not preceded by future-directed intentions. These intentions do not incorporate the results of any prior deliberation, they don’t set the stage for any further planning and they don’t provide a basis for any coordination. The first worry raised by Velleman is thus that these intentions do not seem to serve any of the pragmatic purposes that, on Bratman’s account, constitute the *raison d’être* of intentions. Second, Velleman points out that a similar worry arises for the intentions involved in various cases of planning. He illustrates his point

with a voting example. He argues that while there may be good reasons for my starting to think about my vote in advance, such as giving me sufficient time to deliberate, there doesn’t seem to be any good reason for settling in advance of my arrival in the voting booth whom I will vote for. On the contrary, settling in advance seems to carry potential costs, by making me resistant to reconsideration, without procuring any benefits, since the actual act of casting my ballot doesn’t require any particular prior preparation. Thus, at least in these cases where no further planning is needed once one has settled on a course of action, it is unclear what purpose settling in advance could serve.

Velleman’s third worry relates to Bratman’s view that intention need not imply belief. Bratman indeed maintains that “there need be no irrationality in intending to A and yet still not believing one will”, but that, in contrast, “there will normally be irrationality in intending to A and believing one will not A” (1987, p. 38). According to Velleman, this view of Bratman’s leaves much of his functional account of intentions unmotivated. In particular, it becomes unclear why in intending to A, an agent should be rationally required to identify means of A-ing or to rationally constrain her subsequent practical reasoning by ruling out options inconsistent with her A-ing, if she is agnostic whether she will in fact carry out her intention. Similarly, it becomes unclear why we should impose an agglomerativity requirement on intentions. As Velleman points out, it is unclear why intentions should be jointly executable if the agent can be agnostic as to whether they will be executed.

In my view, Velleman’s third worry is exaggerated. Firstly, while Bratman indeed maintains that an intention to A does not require belief that one will A, he insists at the same time that an intention to A normally supports the belief that one will A. Secondly, Bratman also makes the point that agnosticism about whether one will act as intended does not directly undermine coherent planning but makes it more complex, leading us to form conditional intentions and plans for both failure and success to act as intended. Of course,

the viability of such a move depends on agnosticism being the exception rather than the rule; otherwise, we would have an unmanageable proliferation of conditional branching in our plans.

Velleman's first and second worries run deeper. If the only purposes of intentions are the pragmatic functions Bratman identifies, then there appear to be many instances where intentions don't serve these purposes or where serving them is actually counterproductive. This may be taken to indicate that Bratman's account is incomplete and that he has overlooked some of the functions intentions serve. This line of thought can be pursued in two different directions. On the one hand, we may try to identify further pragmatic functions that intentions, including present-directed intentions, could serve; on the other hand, we may look for non-pragmatic functions that intentions could serve. As we will now see, Velleman explores the second option, turning to Anscombe's theory of intentions in search of an answer. In contrast, what I will do myself later in this paper is explore the first option, giving it a social twist.

Velleman argues that Bratman's account of intentions misses an important function of intention, a function that is a central theme in Anscombe's theory of intention. In her book *Intention* (1963), she argues that intentions provide us with a special kind of self-knowledge and claims that this knowledge is special in two ways. It is knowledge of our own intentional actions, i.e., knowledge not just of what one is attempting to do, but of what one is actually doing, and it is knowledge without observation. Much philosophical ink has been spilled on how exactly these two claims should be interpreted. Following Falvey (2000), Velleman favors a reliabilist interpretation of these claims. According to this interpretation, knowledge of one's own intentional actions is non-observational because it is given by the content of our intentions and intentions in turn normally constitute (practical) knowledge of our own intentional actions because they reliably cause the facts that make them true. Note also, that on this reliabilist reading, Anscombe's claim is not that the content of our intentions provides us with infallible

knowledge of what we are doing. To say that there normally exists a reliable connection between our intentions and actions is not to say that there cannot be cases when this connection does not obtain. However, as Velleman emphasizes, on Anscombe's account, failures of reliability undermine not just the epistemic status of intentions, they also undermine the intentionality of actions. If my intending to *A* does not reliably cause my *A*-ing, then, on the one hand, my intending to *A* will not amount to knowledge that I am *A*-ing and, on the other hand, my *A*-ing when it happens will be an accident rather than an intentional action. According to Anscombe, intentional actions are those "to which the question 'Why?' is given application" (1963, p. 9) and having practical knowledge is knowing a description of what one is doing, has done or is proposing to do that answers the question "Why?". Thus, the basic epistemic function of intentions is to provide us with a form of self-knowledge and self-understanding qua intentional agents.

According to Velleman, acknowledging this epistemic function of intentions does much to alleviate the worries raised by Bratman's practical account. With respect to the first worry – that present-directed intentions serve no purpose – one can now argue that while they might serve no practical purpose they still serve an epistemic function. With respect to the second worry – that on many occasions making one's mind in advance serves no pragmatic purpose –, one can now reply that in matters that are important to one's self-conception, uncertainty about one's future behavior is both uncomfortable and undesirable and that forming an intention allows us to gain self-knowledge and avoid this mental discomfort. With respect to the third worry – that absent a strong enough connection between intention and belief, it is unclear why intentions should be subject to the practical rationality requirements emphasized by Bratman –, Anscombe's theory regarding the epistemic function of intentions lets us see how the epistemic role of intentions could support their pragmatic functions.

The story as told so far suggests that we should think of the epistemic and pragmatic

functions of intentions as complementary. However, as Velleman points out, it still leaves us with two possible hypotheses or creation myths about the origin and ultimate purpose of intentions. On the pragmatic creation myth, the ultimate purpose of intentions would be pragmatic and their epistemic function would be subservient to their pragmatic functions, but may occasionally exemplify re-purposing: “That is, intention might have been designed to embody self-knowledge for the sake of facilitation coordination, but it might then be used on occasion, for the sake of self-knowledge alone, when coordination isn’t necessary” (Velleman 2007, p. 208). By contrast, on the epistemic creation myth, the ultimate purpose of intentions may be to embody self-knowledge, and the pragmatic functions of intentions might have emerged simply as a fortuitous by-product of self-knowledge.

While Velleman has more sympathy for the epistemic than for the pragmatic creation myth, he thinks both should ultimately be rejected. In the next section, I’ll consider his reasons for rejecting them, discuss the alternative story he proposes, and advance my own reasons for being skeptical about this story.

3 Velleman’s spandrel

Despite their differences, the epistemic and the pragmatic creation myths rest on the common assumption that intentions have a teleofunction, some ultimate purpose they are designed to serve. Velleman thinks it is more plausible that their existence is an accident, that is to say, that they are the byproduct of some more general endowments of human nature. In other words, Velleman is tempted to think of the human will as, in Gould & Lewontin’s phrase (1979), a spandrel, a feature formed not by design but as an accidental byproduct of some other designed feature or features. This leads him to be skeptical about both teleological myths. In telling his own creation myth, Velleman pursues two aims. His first aim is to show that the assumption behind the two teleological myths can be dispensed with. His second aim is to show that the accident that led to the emergence of the human will more closely approximates

ates the epistemic than the pragmatic creation myth.

Velleman’s own account of intentions characterizes them as an agent’s commitment to the truth of some act-description of his or her forthcoming behavior that reliably causes this act-description to come true. He argues that this account of intentions “posits nothing more than the predictable consequences of two motivational states whose utility in the design of a creature is far more general than that of the human will” (Velleman 2007, p. 211). In other words, the human will is a spandrel, a feature arising from the accidental confluence of two designed features. What are these two features? The first, according to Velleman, is curiosity, defined as the creature’s drive to understand what goes on in its environment. The second is self-awareness, through which the creature realizes that it is part of its environment and that its own behavior is part of what goes on in this environment. Self-awareness thus allows a creature to acquire an objective conception of itself. A creature that is both curious and self-aware will in turn be driven to understand its own behavior, that is, to understand “how the egocentrically conceived world of doing things is connected to the objectively conceived world of things understood” (Velleman 2007, p. 211). In understanding this, it will have acquired the capacity for intentions.

We can now see why Velleman thinks his own creation myth has more affinities with the epistemic than with the pragmatic creation myth. Curiosity is an epistemic drive and self-awareness is an epistemic capacity. As their byproduct, the capacity for intentions inherits this essential epistemic dimension. We can also understand why he means his own myth as an antidote to the methodological assumption inherent to the idea that intentions serve a specific teleofunction. Curiosity and self-awareness are, Velleman claims, designed for far more general purposes than that of the human will.

I think, however, that this is also where the creation myth told by Velleman reaches its limits. Important questions are left unanswered: What are these more general purposes served by curiosity and self-awareness? What good is curiosity?

What good is self-awareness? Unless he is willing to consider the will as a spandrel of spandrels, Velleman owes us answers to these questions. From an evolutionary point of view, it is unclear what benefits knowledge of their environment and knowledge of themselves could confer on creatures endowed with curiosity and self-awareness unless this knowledge found some behavioral expression. It isn't too difficult to see how a better understanding of their environment can promote more effective behavior, enhance the satisfaction of desires and needs, and ultimately have a differential impact on reproductive success in creatures endowed with curiosity. One should note, however, that pushing Velleman's story one step further in his direction has the effect of undermining his claim that his own myth has strong affinities with the epistemic creation myth for it suggests that the epistemic function of curiosity is ancillary to its pragmatic function, rather than the reverse.

It is less obvious how we should answer the question what good is self-awareness, what purposes it is designed for. My aim in the next two sections will be to remove two obstacles that prevent us from looking in the right direction for an answer to this question. The first obstacle lies in the fact that philosophers have tended to neglect an important pragmatic function of intentions. Thus, Velleman notes, rightly in my view, that Bratman's account of the pragmatic functions of intentions leaves many present-directed intentions without a purpose. However, rather than looking for some further pragmatic purpose intentions may serve, beyond scheduling deliberation and enhancing action coordination over time, Velleman turns his attention to epistemic functions. I will argue in section 4 that they both neglect a further important pragmatic function of intentions, namely their role in the online monitoring and control of action. The second obstacle lies in the fact that one central feature that makes us human, our deep sociality, is either ignored or at best a peripheral concern in philosophical accounts of intentions. Of course, I am not denying the obvious: many philosophers, and Bratman prominently among them, have explored joint agency and collective intentionality. Typically, however, their focus has been on whether or not joint agency should be seen as continuous with individual agency and thus on

whether or not the conceptual framework developed to account for individual intentions could be fruitfully extended to shared intentions.¹ Rarely if ever, however, do they consider the possibility that shared intentions may shed light on some of the features and functions of individual intentions. In section 5, I will argue that the control and monitoring function of intentions plays a crucial role in contexts of joint action. I will further argue that this function might indeed be the primary function of intentions and might have become established because of the role it serves in solving the coordination problems that arise in joint action and because of the benefit thus conferred on creatures capable of solving these coordination problems.

4 Control: A further pragmatic function of intentions

Bratman (1987) considers future-directed intentions as the central case of intending to act and contrasts this approach to intention with an alternative approach that gives priority to immediate intentions or intentions in action. He notes that this second approach naturally leads to the idea that intentions in action reduce to complexes of beliefs and desires, i.e., that what makes it the case that an agent acts with a certain intention are simply facts about the relation between the agent's actions and his beliefs and desires, and that this in turn tempts us into thinking that the same reductive strategy can be extended to future-directed intentions.² Focusing instead on future-directed intentions as the central case of intending allows us to identify functions of intentions that cannot easily be accommodated within a belief-desire model and thus makes the reductive strategy much less appealing. This would account for Bratman's emphasis on the deliberative and coordination functions of intentions. The flip side of the coin, however, is that present-directed intentions are then seen as little more than transmission belts needed to convey the motivational force of future-directed intentions. As noted by

1 See e.g., Bratman (2014) for a positive answer to these questions and Gilbert (1992, 2009) for a negative answer.

2 See for instance Davidson (1980, Essay 1) and Goldman (1970) for belief-desire reductive models of intentions.

Velleman, this leaves us with a potentially large class of actions where present-directed intentions appear to have no role to play, namely all these actions that are intentional yet not preceded by future-directed intentions. What belief-desire reductive approaches, Bratman's account and Velleman's account all seem to overlook is a specific pragmatic function of intentions in action or present-directed intentions, namely their role in the guidance, control and monitoring of action execution.

Harry Frankfurt (1978) was one of the first philosophers to criticize this oversight and insist on the importance of this pragmatic function of intentions. He emphasized that "a person must be in some particular relation to the movements of his body during the period of time in which he is presumed to be performing an action" (Frankfurt 1978, p. 157) and characterized this relation as one of guidance. Other philosophers have since shared his insight. For instance, Brand (1984), Bishop (1989) and Mele (1992) all insist that an adequate account of intentions should incorporate the guiding and monitoring roles of intentions in order to properly capture the close and continuous connection between intention and ongoing action.

The main reason why this connection between intention and ongoing action is needed is that human agents are neither infallible nor omniscient. Their expectations about the circumstances in which the action is to take place may not always be correct and they may fail to anticipate some of the relevant aspects of the situation of action. In other words, their situational beliefs may be incorrect or incomplete. The same goes for their instrumental beliefs. Suppose, for instance, that I intend to visit a colleague in her office. I may be wrong in thinking that this is the door to her office (incorrect situational belief) or unsure which door is her office door (lack of relevant situational belief). Similarly, I may also be wrong in thinking that I should pull the door to open it (incorrect instrumental belief) or unsure whether to push or pull (lack of relevant instrumental belief). If intentions are to reliably produce behavior matching their representational content (e.g., visiting my col-

league in her office), they should have some flexibility and incorporate monitoring processes to detect deviations that jeopardize the success of the action and correction processes to trigger compensatory activity.

This emphasis on control finds a strong echo in the literature on motor cognition (see, e.g., Jeannerod 1997, 2006). Indeed, it is in this literature that we can find the most precise characterization of the monitoring and control functions of intentions and of the mechanisms that support them. According to the very influential internal model theory of motor control, motor control strategies are based on the coupling of two types of internal models: inverse models and forward models (Frith et al. 2000; Jordan & Wolpert 1999; Wolpert 1997). Inverse models compute the motor commands needed for achieving a desired state given the current state of the system and of the environment. An efference copy of these commands is fed to forward models, whose role is to make predictions about the consequences of the execution of these commands. The control of action is thought to depend on the coupling of inverse and forward models through a series of comparators: error signals arising from the comparison of desired, predicted, and actual states (monitoring) are used for various kinds of regulation (control). In particular, they can be used to correct and adjust the ongoing action in the face of perturbations, as well as to update both inverse and forward models to improve their future functioning.

Recent experimental work in motor cognition also suggests, however, that much of action control is automatic and proceeds independently of conscious awareness. For instance, in an experiment (Castiello et al. 1991) participants were asked to reach for and grasp a target as quickly as possible and their hand trajectories were recorded. On some trials, though, the target shifted position after the movement had started. When this happened, participants were instructed to correct their movement in order to reach accurately for the target and to signal the time at which they became aware of its displacement by shouting "Tah!". The experiment showed that the participants started correcting

their movements more than 300ms before they signaled awareness of the target displacement. A subsequent study (Pisella et al. 2000) was especially instructive. In a first experiment they used a similar paradigm but introduced a condition where participants were requested to interrupt their movement when the target changed location. Despite the instruction, the participants could not prevent themselves from correcting their movements instead of stopping for a good 200 ms. In contrast, however, in a second experiment green and red targets were presented simultaneously in the two positions and the participants' task was to point at the green one. On some trials, the color of the two targets could be unexpectedly interchanged at movement onset. When this happened, one group of participants was instructed to interrupt their ongoing movement and the other group to correct it. In contrast to what happened in the first experiment, no automatic corrective movements were observed in the group instructed to interrupt their movement and in the other group corrections involved a significant increase in movement time. Thus, these results suggest that while corrections made in response to spatial perturbations are under automatic control, corrections in response to chromatic perturbations require intentional control.

On the one hand, the mere fact that some or much of action control can be automatic is not a sufficient reason to deny a control function to intentions. The experimental studies presented in the previous paragraph suggest that action control can indeed operate automatically and outside of conscious awareness and that when there is a conflict between automatic and intentional control, automatic control may take precedence over intentional control. Yet, they also provide evidence that some corrections cannot be carried out automatically but depend on intentional control. On the other hand, the mere fact that intentional control seems needed to compensate for chromatic perturbations may not provide sufficient ground for considering that the intentional control of action execution is a central function of intentions. One would want a more systematic account of the respective roles of automatic and intentional con-

trol. Recent developments of the internal model approach to motor control may constitute a useful guide.

While the internal model approach to motor control was initially introduced to account for fine-grained aspects of motor control, more recent versions of this approach emphasize the hierarchical nature of motor control (Hamilton & Grafton 2007; Jeannerod 1997; Kilner et al. 2007). They propose that internal inverse and forward models are arranged in a hierarchy and that error signals generated at one level of the hierarchy can propagate to the next level when correction mechanisms at this level are not able to make the necessary compensations. I have suggested elsewhere (Pacherie 2008) that one can distinguish three broad levels in an action specification hierarchy. At the highest level, action representations represent the whole action as a unit, in terms of its overarching goal and of the sequence of steps or subgoals needed to achieve that goal. At this level, the action may still be represented in a rather abstract format. The second level is concerned with the implementation of each step in the action plan and involves selecting an appropriate motor program given the immediate goal and contextual information about the current state of the agent and the current state of its environment. In other words, processes at this level are in charge of anchoring the successive steps of the action plan in the current situation and of selecting appropriate motor programs. Finally, once a motor program has been selected, the exact values of its parameters must still be set. This is done at the third level, where incoming sensory information about external constraints is used to specify these values.

Acknowledging the existence of different levels of action control corresponding to these different levels in the action specification hierarchy may allow us to accommodate both automatic and intentional action control processes. As long as error signals can be reduced by automatic corrections made at lower levels in the hierarchy, there is no need for the intervention of intentional control. However, there are two classes of cases where automatic corrections may not be sufficient to put an action back on

track. First, important external perturbations can lead to discrepancies that are too large to be automatically compensated. In such a case, error signals would propagate upwards, we would become aware of them and shift to a conscious, intentional compensation strategy. Second, in some instances there may also be discrepancies in the ways the action is or can be specified at different levels of the action representation hierarchy (inter-level representational misalignment). Thus, the study by Pisella and colleagues (Pisella et al. 2000) suggests that action specification at the sensorimotor level does not encode chromatic information and uses spatial information as a proxy for it. When chromatic information and spatial information vary independently, as they do in one of the conditions of the experiment, representations at different levels of the action representation hierarchy become misaligned and the intervention of conscious control becomes necessary to realign them.

Importantly, on this conception of intentional control and as Frankfurt had already noted, what is essential for actions to be intentionally controlled is not that intentional control processes actually affect their course, but that these control mechanisms would have intervened to adjust the action had the need arisen. In other words, an action may be intentionally controlled even though automatic rather than voluntary control mechanisms intervene to compensate for deviations, provided these voluntary control mechanisms would have kicked in, had automatic corrections proved insufficient.

Even more importantly, if action control is an essential function of intentions, then we should stop thinking of intentions as simply mental representations of goals somehow triggering motor processes that, if everything goes well, will yield the desired outcome. Rather, we should think of monitoring and control processes as intrinsic to intentions, that is, of intentions as encompassing not just representations of goals but also a specific set of monitoring and control processes organizing and structuring the motor processes that themselves generate movements.

In this section, I argued for the idea that the control of action execution is an important pragmatic function of intentions. Acknowledging the existence and importance of this function allows us to plug gaps in the creation myths considered earlier. First, it allows us to attribute a specific pragmatic function to present-directed intentions rather than considering them as mere transmission belts in charge of conveying the motivational force of future-directed intentions. We can thus assuage one of the main worries raised by Velleman against Bratman's pragmatic account of intentions and the pragmatic creation myth derived from it. Second, Anscombe's and Velleman's accounts of intentions both assume that intentions reliably cause behavior that matches their representational content. Human agents, however, are neither infallible nor omniscient. Their situational and instrumental beliefs can be incorrect or they can lack situational and instrumental beliefs that are relevant to the successful execution of their intentions. Thus, the reliability demanded by Anscombe's and Velleman's accounts largely depends on our having powerful and flexible control processes allowing us to put our actions back on track when perturbations deviate their course.

One may agree that the conscious control of individual action is a function of intention in the sense that intentions have this causal role, but still be skeptical that this is the role intentions are designed for, or to put it in other words, that it is a teleofunction of intentions. Thus, one could argue that very large external perturbations are rare and that inter-level representational misalignment is the exception rather than the rule. If so, most of action control would be automatic anyway and intentional action control would play at best a marginal role. It would therefore be unlikely to confer on intention-forming creatures benefits important enough to warrant the claim that intentions are designed for action control. As I have tried to argue in this section, the benefits conferred by online conscious control over actions are not as negligible as this deflationary view implies. In addition, I think we can build a very strong case that conscious action control confers im-

portant benefits if we consider joint activities rather than just individual actions. Acting jointly demands that we solve coordination problems that do not arise (or arise only in a very attenuated form) in individual action. In what follows, I will argue that online conscious control plays a crucial role in solving these coordination problems. I will further speculate that conscious online control over actions might indeed have become established as the primary function of intentions because of the role it served in solving these coordination problems and because of the benefit this conferred on creatures capable of solving these coordination problems and thus of acting jointly in an efficient and flexible way.

5 The social creation myth

Humans have been characterized as the ultra-cooperative species (Tomasello 2009, 2011). This ultra-cooperativeness has made us one of the most successful species on earth, spreading all over the planet, creating and developing cultural artifacts and practices that are themselves culturally transmitted and accumulate over time, thus giving us a further competitive edge over other species. According to Tomasello, underlying humans' ultra-cooperativeness are a set of species-unique skills and motivations for shared intentionality, involving "such things as the ability and motivation to form shared goals and intentions with others in collaborative activities, and the ability and motivation to share experience with others via joint attention, cooperative communication, and teaching" (2011, p. 6).

The gist of the social creation myth I am proposing in this section is that the main benefits associated with intentions and with the kind of control over actions they make possible arise in social cooperative contexts where agents have to coordinate their actions to achieve a shared goal. I start with an examination of the special demands for coordination acting jointly with others creates. I then explain how the capacity to form conscious intentions is a crucial component of our ability to meet these demands.

Successful joint action depends on the efficient coordination of participant agents' goals, intentions, plans, and actions. As I argued elsewhere (Pacherie 2012), it is not enough that agents control their own actions, i.e., correctly predict their effects, monitor their execution and make adjustments if needed. They must also coordinate their actions with those of their co-agents so as to achieve their joint goal. For that they must monitor their partner's intentions and actions, predict their expected consequences and use these predictions to adjust what they are doing to what their partners are doing. The implication of these processes, however, is not unique to joint action nor enough to promote their success. In competitive contexts they also play an important role. For instance, in a fight being able to anticipate your opponent's moves and to act accordingly is also crucial. What is furthermore required in the case of joint action is that co-agents share a goal and understand the combined impact of their respective intentions and actions on their joint goal and adjust them accordingly. In competitive contexts, an agent should typically aim at predicting his opponents' moves, while at the same time endeavoring to make his own moves unpredictable to his opponents. In contrast, in cooperative contexts mutual predictability must be achieved for efficient coordination towards a shared goal to be possible. Agents should be able to align their representations of what themselves and their partners are doing and of how these actions together contribute to the shared goal.

Various forms of uncertainty can undermine mutual predictability, the alignment of representations and hence coordination. They can be organized into three broad categories. The first category involves motivational uncertainty: we can be unsure how convergent a potential partner's interests are with our own interests and thus unsure whether there are goals we share and can promote together. The second category involves instrumental uncertainty: even assuming that we share a goal, we can be unsure what we should do to achieve that goal, or, if we have a plan, unsure how roles should be distributed among us, or, yet, unsure when and where we should act. The third category involves common ground uncertainty: we can be

unsure how much of what is relevant to our deciding on a joint goal, planning for that goal and executing our plan is common ground or mutually manifest to us.

Philosophical accounts of joint agency, including Bratman's (2009, 2014) do not ignore these challenges but they are essentially concerned with high-level processes involved in making decisions about whether or not to act together and in advance planning. Their focus is on the coordination of agent's intentions prior to acting and they pay little heed to the processes enabling people to coordinate during action execution. In contrast, in the last decade, cognitive scientists have investigated joint action by focusing on lower-level online coordination processes in relatively simple joint tasks and on the factors that affect these coordination processes. In what follows, I will argue that there are important limitations to what these advance and online coordination processes can achieve and that high-level online intentional control is crucial to overcoming these limitations. First, however, let us consider the main characteristics of the two sets of coordination processes philosophers and psychologists typically focus on.

Bratman's account of shared intentions is a good illustration of the way philosophical accounts approach coordination issues in joint action. In addition, its explicitness makes it possible to see clearly what advance coordination involves and how it is achieved.

Bratman (2009) proposes that shared intention involves the following conditions as its main building blocks:

1. Intentions on the part of each in favor of the joint activity.
2. Interlocking intentions: each intends that the joint activity go in part by way of the relevant intentions of each of the participants.
3. Intentions in favor of meshing subplans: each intends that the joint activity proceed by way of subplans of the participants that are co-realizable and can be consistently agglomerated.
4. Disposition to help if needed: given that the contribution of the other participants to the

joint activity is part of what each intends, and given the demands of means-end coherence and of consistency that apply to intentions, each is under rational pressure to help others fulfill their role if needed.

5. Interdependence in the persistence of each participant's relevant intention: each believes that the persistence of the other participants' intention in favor of the joint activity depends on the persistence of his own and vice-versa.
6. Joint-action-tracking mutual responsiveness: each is responsive to each in relevant subsidiary intentions and in relevant actions in a way that tracks the joint action.
7. Common knowledge among all participants of all these conditions.

Let me offer some comments on these conditions. First, Bratman offers these conditions as a set of sufficient conditions for a shared intention, leaving it open that shared intentions may be realized in other ways, in particular in cases of joint activities involving institutions. Second, conditions (1), (2) and (5) are meant to deal with motivational uncertainty. Bratman points out that the concept of a joint activity that figures in the contents of the intentions in (1) should be understood in a way that is neutral with respect to shared intentionality. So condition (1) only insures that agents share goals in a weak sense of the notion. Rather it is condition (2) that is in charge of insuring that the motivational states of the agents align in the way required for joint cooperative activity: it is the fact that for each participant, the content of their intention refers to the role of the intentions of other participants that, for Bratman, captures the intentional jointness of their actions. Condition (5) in turn specifies how these motivations stay aligned. Third, conditions (3), (4) and (6) relate to means-end uncertainty and are meant to reduce it. According to Bratman, they can be derived from condition (2) taken together with the norms of practical rationality that already govern individual planning and acting. Bratman's key idea is that the interlocking and interdependent intentions of individual participants, in responding to the norms of

practical rationality governing individual planning agency, will also respond to the norms of social agglomeration and consistency, social coherence and social stability shared intentions are subject to. This would involve, in Bratman's terms, commitments to mutual compatibility of relevant sub-plans, commitments to mutual support, and joint-action tracking mutual responsiveness. Finally, the function of condition (7) is, rather obviously, to reduce common ground uncertainty.

Bratman's basic idea is thus that this structure of interlocking and interdependent intentions, when it functions properly, frames relevant bargaining and shared deliberation and thus supports and guides coordinated planning and action in pursuit of the intended shared activity. Unsurprisingly, since Bratman's theory of joint agency is continuous with his planning theory of individual intentions, it is in virtue of the pragmatic functions intentions already serve in the individual action case that the interlocking and interdependent intentions of individual participants can also support coordination in the joint action case.

While Bratman, in his condition (6), stipulates that agents should be mutually responsive not just in their relevant intentions and subsidiary intentions but also in relevant actions in a way that tracks the joint action, his account doesn't tell us by what means mutual responsiveness in action is achieved. To know more about this, we have to turn our attention to recent psychological work on joint agency. In contrast to philosophical approaches, cognitive psychology studies of joint action typically focus on the perceptual, cognitive, and motor processes that enable individuals to coordinate their actions with others online.

Knoblich and colleagues (Knoblich et al. 2011) distinguish between two broad categories of coordination processes, emergent and planned. In emergent coordination, coordinated behavior occurs due to perception-action couplings that make multiple individuals act in similar ways. One source of emergent coordination is entrainment, the process of synchronizing two or more actors' rhythmic behaviors with respect to phase (e.g., Richardson et al. 2007). A second

source of emergent coordination is perception-action matching, whereby observed actions are matched onto the observer's own action repertoire and can induce the same action tendencies in different agents who observe one another's actions (Jeannerod 1999; Prinz 1997; Rizzolatti & Sinigaglia 2010; Knoblich & Sebanz 2008). Importantly, emergent forms of coordination are independent of any joint plans or common knowledge, which may be altogether absent. They support basic forms of motor and representational alignment that can facilitate mutual responsiveness in action, but they do not ensure that the agent's actions track a joint goal. Indeed, the successful performance of some joint actions may require that these automatic coordination processes be inhibited. For instance, the performance of composer Steve Reich's famous piece, *Drumming*, based on the technique of phasing, requires the musicians to play the same rhythmic pattern out of sync.

In planned coordination, agents plan their own actions in relation to the joint goal and also to some extent to their partners' actions. As emphasized by Knoblich et al. (2011), shared task representations play an important role in planned coordination. Shared task representations do not only specify in advance what the respective tasks of each of the co-agents are, they also provide control structures that allow agents to monitor and predict what their partners are doing, thus enabling interpersonal coordination in real time. Empirical evidence shows that having shared task representations influences perceptual information processing, action monitoring, control and prediction during the ensuing interaction (Heed et al. 2010; Schuch & Tipper 2007; Sebanz et al. 2006; Tsai et al. 2006). Furthermore, several studies (Sebanz et al. 2005; Sebanz et al. 2006) have shown that actors may form shared representations of tasks quasi-automatically, even when it is more effective to ignore one another.

Several researchers have also suggested that joint attention provides a basic mechanism for sharing representations of objects and events and thus for creating a perceptual common ground in joint action (Tomasello & Carpenter 2007; Tollefsen 2005). To act jointly, it is often

necessary not only that the co-agents identify the objects to be acted upon, their location as well as the location of possible obstacles, but also be mutually aware that they do. Joint attention may thus play an important role in ensuring that co-agents track the same objects and features of the situation and be mutually aware that they do. In a recent study, Böckler et al. (2011) showed that attending to objects together from opposite perspectives makes people adopt an allocentric rather than the default egocentric frame of reference. These authors suggest that taking an allocentric reference may support the efficiency of joint actions from different spatial orientations. Independently of mutual manifestness, being able to assess what others are perceiving, or can or cannot perceive at a given moment in time may also facilitate coordination. For instance, a study by Brennan and colleagues (Brennan et al. 2007) demonstrated that co-agents in joint visual search space were able to distribute a common space between them by directing their attention depending on where the other was looking and that their joint search performance was thus much more efficient than their performance in an individual version of the search task.

There are, however, important limitations to what these emergent and planned on-line coordination processes can achieve. First, to the extent that they exploit perceptual information, they can be of no help unless a certain amount of common perceptual information is indeed available to co-agents. Second, even when common perceptual information is available, there are limits to our processing capacities. An agent may be able to simultaneously track what a small number of other agents are currently doing or attending to, but when the number of agents increases, this capacity soon finds its limits. Our capacity to co-represent the actions, goals, and intentions of other agents we observe acting encounters similar limitations. Understanding of actions through motor resonance and mirroring works only to the extent that the observed actions are part of the action repertoire of the observer. Similarly, when actions are relatively

novel, agents may not yet have formed sufficiently detailed shared task representations. Finally, unexpected effects of action execution or failures of coordination may reveal various forms of misalignment between partners' representations or indicate that their representations, though aligned, were inaccurate.

When pre-alignment is insufficient or breakdowns occur due to misalignment in the action execution phase, the deliberate and conscious production of social signals aimed at aligning or realigning relevant representations becomes crucial. Agents cannot count on alignment arising spontaneously. They have to make it happen. Intentional communication, whether verbal or not, is then needed to make it happen.

As emphasized by Herbert Clark (2006), joint activities can typically be partitioned into two sets of actions: a basic joint activity and coordinating joint actions. The basic joint activity comprises all the actions essential to achieving the basic joint goal, while the coordinating joint actions consists in the set of communicative acts about the basic activities that insure relevant representational alignment. To study this partitioning of joint activities, Clark and his co-workers ushered two people in a small room, gave them the parts of a kit for a TV stand and asked them to assemble the stand, videotaping them and recording their verbal exchanges while putting they did it. Here's a short extract of their exchanges, taken from Clark (2006, p. 128):

Ann Should we put this in, this, this little like kinda cross bar, like the T? like the I bar?

Burton Yeah ((we can do that))

Ann So, you wanna stick the ((screws in)). Or wait is, is, are these these things, or?

Burton That's these things I bet. Because there's no screws.

Ann Yeah, you're right. Yeah, probably. If they'll stay in.

Burton I don't know how they'll stay in ((but))

Ann Right there.

Burton Is this one big enough?

Ann Oh ((xxx)) I guess cause like there's no other side for it to come out.

Burton M-hm.

[8.15 sec]

Burton ((Now let's do this one))

Ann Okay

First, it should be noted that, as often happens in daily life, this joint activity was not planned in advance. Instead, Ann and Burton discover that they have to assemble a TV stand and work out together what they should do as they go along. Second, Clark points out that Ann and Burton's coordinating joint actions are structured in what he calls projective pairs, comprising a proposal and an uptake (i.e., full acceptance, altered acceptance or rejection of proposal). Third, the exchanges can be gestural as well as verbal. For instance, instead of, or concomitantly with, asking verbally whether Burton is ready to fasten the screws, Ann may present him with the screwdriver and his taking it count as acceptance. Fourth, the contents of these exchanges show that they are aimed at reducing instrumental uncertainty. Typically, they are about what should be done and how, who should do what, and when and where it should be done. When the task presents difficulties, they may also serve to reduce motivational uncertainty. For instance, Burton might ask whether they should give it a last try and Ann either acquiesce or reject the proposal. Finally, the structure of the projective pairs shows that at the same time they aim at reducing common ground uncertainty. Proposals are about potential alignments and full acceptance confirms alignment and common ground. Tellingly, with altered acceptance uptakes, projective pairs evolve into projective triads, the third element of the exchange being the proposer's uptake on the alteration.

Importantly, to negotiate and achieve alignment in this way, we must be aware of our own and others' intentions and beliefs and this at two levels, corresponding to the two sides of the partitioning characterized by Clark. On the one hand, it is essential to the fulfillment of communicative intentions that they be recognized as such by the addressee (Grice 1957; Recanati 1986; Sperber & Wilson 1986). On the other hand, what agents communicate in these contexts is information about their beliefs and intentions regarding the joint action. This suggests that the development of self-consciousness and consciousness of other minds, of intentional communication, and of increasingly complex forms of coordinated joint action go hand in hand.

The success of both individual and joint action depends on representational alignment. In the case of individual action, representation alignment takes two main forms. First, at a given level of action specification, a match should be achieved between representations of desired, predicted and actual states. We can call this first form of alignment intra-level alignment. Second, inter-level alignment is also necessary; that is, despite differences in representational format and resources, action specifications at different levels of the action representation hierarchy should be kept aligned. Conscious online control may be needed to restore alignment when severe intra- or inter-level discrepancies occur. However, it may be argued that in the individual case alignments are taking place within a single cognitive system and that this system is normally sufficiently integrated or unified that serious misalignments are rare and thus that the need for online conscious control is limited.

The main difference between individual and joint actions lies in the coordination demands essential to joint action. Thus, a third form of representational alignment becomes crucial in joint action. In addition to individual intra- and inter-level representational alignment, inter-agent representational alignment is necessary to meet coordination demands. Inter-agent alignment may be achieved in part through advanced planning, as proposed by Bratman. It

can also be achieved in part through online emergent and planned coordination processes of the types explored and described in the recent psychological literature. However, there are important limitations to what these coordination processes can achieve. Advance planning, when it takes place, may help define a shared background framework for the joint action, but at this stage it is typically impossible to anticipate all the coordination demands that will arise at the execution stage. Some of these demands may be met by the kinds of online coordination processes reviewed earlier in this section, but, as I pointed out, there are also important limitations to what they can achieve. In many instances, the progress of a joint action is hindered or the action breaks down due to various forms of misalignment between agents' representations. In such instances, individual corrections do not suffice to put the joint action back on track. Rather, to overcome these failures, agents need to align or realign their representations. This process calls for what Clark calls coordinating joint actions, that is, communicative acts about the basic joint activity. These communicative acts in turn are intentional and aim at communicating information about the agents' intentions and beliefs with a view to achieve alignment. But one can only communicate intentionally about one's beliefs and intentions if one is aware of them. Conversely, one can only understand the communicative acts of other agents if one realizes that these agents have a capacity for intentions. Finally and crucially, as already emphasized by Velleman (2007) in his discussion of Bratman's account, intentions could not serve their pragmatic functions unless they also had an epistemic role. In other words, if my having the intention to *A* didn't count as a form of practical self-knowledge and didn't give me grounds to believe that I would act as intended, my communicating (sincerely) about my intention to *A* would not license other agents to form beliefs about my future actions and thus would not yield the kind of inter-agent representational alignment needed to achieve coordination.

To recap, joint actions create more comprehensive demands for representational align-

ment than individual actions, since their success depends not just on individual intra- and inter-level representational alignment but also on inter-agent representational alignment. New resources are needed to meet these demands. On the social creation myth proposed here, a capacity for conscious intentions is crucial to inter-agent representational alignment. Having conscious intentions allows us to communicate about them and engage in coordinating joint actions that create common ground and promote the success of basic joint activity. The answer this myth offers to the question what is the purpose of conscious intentions is then that it is to enable more efficient inter-personal coordination in joint action and thus reap the benefits that come with increasingly complex and flexible forms of coordinated actions. The social creation myth doesn't deny intentions an epistemic role. On the contrary, it acknowledges that intentions couldn't serve their inter-personal coordination function if they did not at the same time provide us with a form of self-knowledge. However, it views their epistemic function as subservient to their coordination function. The social creation myth does not deny either that conscious intentions play a role in the online control of individual action. Rather, it proposes that conscious control of individual action may be a by-product of a capacity for conscious control that became established in social contexts because of the role it served in solving inter-agent coordination problems and because of the benefit conferred by the forms of cooperation it made possible.

6 Conclusion: Relating creation myths

The Bratmanian creation myth is pragmatic but also diachronic and individualist. Intentions have a purpose or teleofunction. This function is pragmatic insofar as the main benefit attached to intentions is to allow us to secure greater desire satisfaction. The way intentions secure this benefit is by allowing us to organize and coordinate our actions diachronically, in other words to become planning agents. As noted by Velleman, this emphasis of diachronicity and future-directed intentions leaves present-directed

intentions without a clear function. Finally, this myth is to a large extent individualist. While planning agency also enables inter-individual coordination, the social dimension of intentions remains secondary in Bratman's account and again his main concern is with diachronically organized joint actions.

While the social creation myth also sees intentions as having a pragmatic purpose, in contrast to the Bratmanian myth, it emphasizes the social and synchronous dimension of intentions. Instead of self-coordination over time, it emphasizes cooperation and flexibly coordinated joint action as the main route to greater desire satisfaction. It thus reverses the Bratmanian perspective in proposing that intentions are designed to enable a more efficient online coordination of joint action and in considering future-directed individual or joint planning as derivative or secondary functions of intentions.

Because its main emphasis is on synchronicity rather than diachronicity, the social creation myth has no problem attributing a pragmatic control function to present-directed intentions. It is thus impervious to one of the attractions of the Anscombian creation myth. We need feel no temptation to attribute an epistemic function to present-directed intentions for lack of any other plausible option. The social creation myth, however, does not dispense with epistemic functions altogether, quite the reverse. Not only is the fact that intentions embody a form of self-knowledge essential to their role in the coordination of joint actions, but in addition the way intentions play their coordinative role is by contributing to the alignment of representations with co-agents and thus to the production of shared knowledge. Thus, on the social creation myth, the epistemic function of intentions is not just to provide us with self-knowledge about our intentions and actions, it is also to contribute to the formation of shared knowledge. However, the social creation myth remains closer to the pragmatic than to the epistemic creation myth in considering that the epistemic function of intentions is ancillary to its pragmatic purpose.

Finally, is the social creation myth a teleological myth or, like Velleman's myth, the

story of a spandrel? I must admit that I am not sure what the answer to this question is or should be. Indeed, this was one of the reasons why I chose to call my story a creation myth. One thing is sure though, if it is a story about a spandrel, this spandrel is not the same as Velleman's. His spandrel is a by-product of curiosity and self-awareness. This spandrel, if it is one, would involve a third element, sociality or cooperativeness. Social theories of consciousness (Frith 2010; Graziano & Kastner 2011) propose that consciousness has evolved to facilitate social interactions and enhance social cooperation. On the one hand, a capacity for consciousness is of course a much more general capacity than a capacity for conscious intentions and this may suggest that the latter, as a by-product of this more general capacity, is itself merely a spandrel. On the other hand, if the ultimate purpose of consciousness is to enhance social cooperation, then conscious intentions are a key element in making this possible and calling our capacity for intention a spandrel would fail to do justice to their role.

References

- Anscombe, G. E. M. (1963). *Intention*. Oxford, UK: Blackwell.
- Bishop, J. C. (1989). *Natural agency: An essay on the causal theory of action*. Cambridge, UK: Cambridge University Press.
- Brand, M. (1984). *Intending and acting: Toward a naturalized action theory*. Cambridge, MA: MIT Press.
- Bratman, M. (1987). *Intention, plans, and practical reason*. Cambridge, MA: Harvard University Press.
- (2009). Modest sociality and the distinctiveness of intention. *Philosophical Studies*, 144 (1), 149-165. [10.1007/s11098-009-9375-9](https://doi.org/10.1007/s11098-009-9375-9)
- (2014). *Shared agency*. Oxford, UK: Oxford University Press.
- Brennan, S. E., Chen, X., Dickinson, C., Neider, M. & Zelinsky, G. (2007). Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition*, 106 (3), 1465-1477. [10.1016/j.cognition.2007.05.012](https://doi.org/10.1016/j.cognition.2007.05.012)
- Böckler, A., Knoblich, G. & Sebanz, N. (2011). Giving a helping hand: Effects of joint attention on mental rotation of body parts. *Experimental Brain Research*, 211 (3-4), 531-545. [10.1007/s00221-011-2625-z](https://doi.org/10.1007/s00221-011-2625-z)
- Castiello, U., Paulignan, Y. & Jeannerod, M. (1991). Temporal dissociation of motor responses and subjective awareness a study in normal subjects. *Brain*, 114 (6), 2639-2655. [10.1093/brain/114.6.2639](https://doi.org/10.1093/brain/114.6.2639)
- Clark, H. H. (2006). Social actions, social commitments. In N. J. Enfield & S. C. Levinson (Eds.) *Roots of human sociality: Culture, cognition, and interaction* (pp. 126-150). Oxford, UK: Berg.
- Davidson, D. (1980). *Essays on Actions and Events*. Oxford, UK: Oxford University Press.
- Falvey, K. (2000). Knowledge in intention. *Philosophical Studies*, 99 (1), 21-44. [10.1023/a:1018775307559](https://doi.org/10.1023/a:1018775307559)
- Frankfurt, H. (1978). The problem of action. *American Philosophical Quarterly*, 15 (2), 157-162.
- Frith, C. (2010). What is consciousness for? *Pragmatics & Cognition*, 18 (3), 497-551. [10.1075/pc.18.3.03fri](https://doi.org/10.1075/pc.18.3.03fri)
- Frith, C. D., Blakemore, S.-J. & Wolpert, D. M. (2000). Abnormalities in the awareness and control of action. *Philosophical Transactions of the Royal Society of London*, 355 (1404), 1771-1788. [10.1098/rstb.2000.0734](https://doi.org/10.1098/rstb.2000.0734)
- Gilbert, M. (1992). *On social facts*. Princeton, NJ: Princeton University Press.
- (2009). Shared intention and personal intentions. *Philosophical Studies*, 144 (1), 167-187. [10.1007/s11098-009-9372-z](https://doi.org/10.1007/s11098-009-9372-z)
- Goldman, A. (1970). *A theory of human action*. Englewood Cliffs, NJ: Prentice-Hall.
- Gould, S. J. & Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society B*, 205 (1161), 581-598. [10.1098/rspb.1979.0086](https://doi.org/10.1098/rspb.1979.0086)
- Graziano, M. S. & Kastner, S. (2011). Human consciousness and its relationship to social neuroscience: A novel hypothesis. *Cognitive Neuroscience*, 2 (2), 98-113. [10.1080/17588928.2011.565121](https://doi.org/10.1080/17588928.2011.565121)
- Grice, H. P. (1957). Meaning. *Philosophical Review*, 66 (3), 377-388. [10.2307/2182440](https://doi.org/10.2307/2182440)
- Hamilton, A. F. & Grafton, S. T. (2007). The motor hierarchy: From kinematics to goals and intentions. In P. Haggard, Y. Rossetti & M. Kawato (Eds.) *Sensorimotor foundations of higher cognition* (pp. 381-408). Oxford, UK: Oxford University Press.
- Heed, T., Habets, B., Sebanz, N. & Knoblich, G. (2010). Others' actions reduce crossmodal integration in peripersonal space. *Current Biology*, 20 (15), 1345-1349. [10.1016/j.cub.2010.05.068](https://doi.org/10.1016/j.cub.2010.05.068)
- Jeannerod, M. (1997). *The cognitive neuroscience of action*. Oxford, UK: Blackwell.
- (1999). The 25th Bartlett Lecture. To act or not to act: Perspectives on the representation of actions. *Quarterly Journal of Experimental Psychology*, 52 (3), 1-29. [10.1080/713755803](https://doi.org/10.1080/713755803)
- (2006). *Motor cognition*. Oxford, UK: Oxford University Press.
- Jordan, M. I. & Wolpert, D. M. (1999). Computational motor control. *The cognitive neurosciences* (pp. 485-493). Cambridge, MA: MIT Press.
- Kilner, J. M., Friston, K. J. & Frith, C. D. (2007). Predictive coding: An account of the mirror neuron system. *Cognitive Processing*, 8 (3), 159-166. [10.1007/s10339-007-0170-2](https://doi.org/10.1007/s10339-007-0170-2)
- Knoblich, G., Butterfill, S. & Sebanz, N. (2011). Psychological research on joint action: Theory and data. *Psychology of Learning and Motivation - Advances in Research and Theory*, 54, 59-101. [10.1016/B978-0-12-385527-5.00003-6](https://doi.org/10.1016/B978-0-12-385527-5.00003-6)
- Knoblich, G. & Sebanz, N. (2008). Evolving intentions for social interaction: From entrainment to joint action. *Philosophical Transactions of the Royal Society B*, 363 (1499), 2021-2031. [10.1098/rstb.2008.0006](https://doi.org/10.1098/rstb.2008.0006)
- Mele, A. R. (1992). *Springs of action: Understanding intentional behavior*. Oxford, UK: Oxford University Press.

- Pacherie, E. (2008). The phenomenology of action: A conceptual framework. *Cognition*, 107 (1), 179-217. [10.1016/j.cognition.2007.09.003](https://doi.org/10.1016/j.cognition.2007.09.003)
- (2012). The phenomenology of joint action: Self-agency vs. joint-agency. In A. Seemann (Ed.) *Joint attention: New developments* (pp. 343-389). Cambridge MA: MIT Press.
- Pisella, L., Grea, H., Tilikete, C., Vighetto, A., Desmurget, M., Rode, G. & Rossetti, Y. (2000). An ‘automatic pilot’ for the hand in human posterior parietal cortex: Toward reinterpreting optic ataxia. *Nature Neuroscience*, 3 (7), 729-736. [10.3389/fnhum.2013.00336](https://doi.org/10.3389/fnhum.2013.00336)
- Prinz, W. (1997). Perception and action planning. *European Journal of Cognitive Psychology*, 9 (2), 129-154. [10.1080/713752551](https://doi.org/10.1080/713752551)
- Recanati, F. (1986). On defining communicative Intentions. *Mind and Language*, 1 (3), 213-242. [10.1111/j.1468-0017.1986.tb00102.x](https://doi.org/10.1111/j.1468-0017.1986.tb00102.x)
- Richardson, M. J., Marsh, K. L., Isenhower, R. W., Goodman, J. R. L. & Schmidt, R. C. (2007). Rocking together: Dynamics of unintentional and intentional interpersonal coordination. *Human Movement Science*, 26 (6), 867-891. [10.1016/j.humov.2007.07.002](https://doi.org/10.1016/j.humov.2007.07.002)
- Rizzolatti, G. & Sinigaglia, C. (2010). The functional role of the parieto-frontal mirror circuit: Interpretations and misinterpretations. *Nature Reviews Neuroscience*, 11 (4), 264-274. [10.1038/nrn2805](https://doi.org/10.1038/nrn2805)
- Schuch, S. & Tipper, S. P. (2007). On observing another person’s actions: Influences of observed inhibition and errors. *Perception & Psychophysics*, 69 (5), 828-837. [10.3758/BF03193782](https://doi.org/10.3758/BF03193782)
- Sebanz, N., Knoblich, G. & Prinz, W. (2005). How two share a task: Corepresenting stimulus-response mappings. *Journal of Experimental Psychology: Human Perception and Performance*, 31 (6), 1234-1246. [10.1037/0096-1523.31.6.1234](https://doi.org/10.1037/0096-1523.31.6.1234)
- Sebanz, N., Knoblich, G., Prinz, W. & Wascher, E. (2006). Twin Peaks: An ERP study of action planning and control in co-acting individuals. *Journal of Cognitive Neuroscience*, 18 (5), 859-870. [10.1162/jocn.2006.18.5.859](https://doi.org/10.1162/jocn.2006.18.5.859)
- Sperber, D. & Wilson, D. (1986). *Relevance: Communication and cognition*. Oxford, UK: Blackwell.
- Tollefsen, D. (2005). Let’s pretend: Children and joint action. *Philosophy of the Social Sciences*, 35 (75), 74-97. [10.1177/0048393104271925](https://doi.org/10.1177/0048393104271925)
- Tomasello, M. (2009). *Why we cooperate*. Cambridge, MA: MIT Press.
- (2011). Human culture in evolutionary perspective. In M. Gelfand (Ed.) *Advances in Culture and Psychology* (pp. 5-51). Oxford, UK: Oxford University Press.
- Tomasello, M. & Carpenter, M. (2007). Shared intentionality. *Developmental Science*, 10 (1), 121-125. [10.1111/j.1467-7687.2007.00573.x](https://doi.org/10.1111/j.1467-7687.2007.00573.x)
- Tsai, C.-C., Kuo, W.-J., Jing, J.-T., Hung, D. L. & Tzeng, O. J.-L. (2006). A common coding framework in self-other interaction: Evidence from joint action task. *Experimental Brain Research*, 175 (2), 353-362. [10.1007/s00221-006-0557-9](https://doi.org/10.1007/s00221-006-0557-9)
- Velleman, D. (2007). What good is a will? In A. Leist & H. Baumann (Eds.) *Action in context* (pp. 193-215). Berlin, GER: de Gruyter.
- Wolpert, D. M. (1997). Computational approaches to motor control. *Trends in Cognitive Sciences*, 1 (6), 209-216. [10.1016/S1364-6613\(97\)01070-X](https://doi.org/10.1016/S1364-6613(97)01070-X)

Conscious Intentions: Do We Need a Creation Myth?

A Commentary on Elisabeth Pacherie

Andrea R. Dreßing

We experience ourselves as agents, performing goal-directed actions in the world. In her paper about *Conscious Intentions: The social creation myth* Pacherie develops a creation myth about the function of conscious intentions, based on her hierarchical concept of individual motor actions and joint action. In this creation myth, conscious intentions are not understood as internal mental states with a teleo-functional role. Having a conscious intention exerts a specific contribution to motor control and conscious intentions might have a potential causal power in this myth.

In this commentary I want to postulate, that Pacherie's social creation myth is more than a myth but rather the search for an explanation of the function of conscious intentions in the physical world. It tries to explain the feature of the intention *being conscious* that endows it with its particular causal function. Yet — speaking about a causal function — the potential analytical and neuroscientific limitations of a causal function of conscious intentions in the social creation myth have to be analysed with regard to the argument of causal closure and results of experimental approaches to the causal relevance of conscious intentions. I argue that despite these limitations the social creation myth could be an important step on the way of finding an explanation about the function of conscious intentions, if the question about the *function* of conscious intentions is slightly adjusted and is not understood in a strictly causal way.

Keywords

Causal closure | Conscious agents | Conscious intention | Creation myth | Intentional action | Joint action | Mental causation | Neuronal correlates of conscious intention

1 Introduction

We experience ourselves as agents, performing goal-directed actions in the world. This can be a short-term goal of a motor action like grasping a glass of water, or long-term goal, like the plan to call someone later on. One crucial point in both cases is that we know what we do or want to do. We are aware of our goals before and during acting. This awareness constitutes a conscious intention to act. Even further, we seem to

control our actions — at least most of the time — through our intentions. We also have a sense of agency for our actions, which is an immediate feeling of control and authorship (Gallagher 2005). Common sense teaches us that consciousness of our intentions seems to be of unquestionable relevance for our everyday acting.

This experience raises two kinds of questions: *Why* do we experience our intentions as

Commentator

Andrea R. Dreßing

andrea.dressing@uniklinik-freiburg.de

Klinik für Neurologie und
Neurophysiologie
Universitätsklinikum Freiburg
Freiburg, Germany

Target Author

Elisabeth Pacherie

elisabeth.pacherie@ens.fr
Ecole Normale Supérieure
Paris, France

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

Table 1: Overview over the different approaches to the explanation of the function of conscious intentions

Anscombe 1963	
Epistemic creation myth	Conscious intentions “provide us with a special kind of self-knowledge” (Pacherie this collection , p. 5)
Bratman 1987	
Pragmatic creation myth	Conscious intentions “[turn us into] temporally extended agents” (Pacherie this collection , p. 3)
Velleman 2007	
Conscious intentions as a spandrel	Conscious intentions are a “by-product of some more general endowments of human nature” (Pacherie this collection , p. 6)
Pacherie this collection	
Social creation myth	conscious intentions “[..are] not just representations of goals but also [...] a specific set of monitoring and control processes, organizing and structuring motor processes that themselves generate movements” (Pacherie this collection , p. 10)

conscious? What is the function of the phenomenal experience of conscious intentions and *how* do intentions exert their role in our acting? These questions address the problem of conscious intentions at two levels. One is about identifying the function and benefits of conscious intentions for our human nature — it is about a myth. The other seems to be above that about understanding, having to do with how the conscious intention exerts its function. It is an attempt to find a scientific, mechanistic explanation about the function of conscious intentions in not only analytical, but also empirical terms (see also [Anderson this collection](#), and [Craver this collection](#)).

In her target article, [Conscious Intentions: The social creation myth](#), Elisabeth Pacherie wants to elucidate the function of conscious intentions and reviews teleological approaches on the role of conscious intentions offered by Velleman, as well as his interpretations of Bratman and Anscombe. In addition, she addresses above-mentioned question about the “how” of the causal role of intentions. Based on her hierarchical concept of individual motor actions and scientific data about joint action, Pacherie develops her own approach to the function of conscious intentions. Her idea is supported by the consideration of the potentially striking role of conscious intentions in joint actions (inter-indi-

vidual actions) regarded as one of the major achievements of the human species. Pacherie’s idea is that conscious intentions have the function of controlling motor action and to intra- and inter-individually align our actions with each other.

Answering the initial question of whether we need a creation myth or not, I would like to answer: no, we do not need a myth. We need, as Pacherie tries to give in her target paper, an explanation. What I perhaps like best about the paper is her focus on the role of conscious intentions in action, while the other creation myths described in her paper only consider a more abstract level. We experience the function of conscious intentions strongly and immediately in individual and joint action. Understanding the function of conscious intentions in this context might therefore be one of the most difficult but promising approaches, as it is so essential for human existence. Her social creation myth has the aim to find an explanation of the function and potential causal role of conscious intentions. The importance of this approach, to my mind, is strengthened by Pacherie’s attempt to combine empirical data and analytical considerations about motor action and motor control.

In what follows, the teleological and social creation myths are first summarized. Postulating that Pacherie’s social creation myth is more



Figure 1: Pacherie’s model of intentional action.

than a myth, it should nevertheless fit the current philosophical conceptions and empirical knowledge about the nature of conscious intentions and their causal function. I therefore analyse it according to contemporary approaches in philosophy of mind and I incorporate knowledge of experimental approaches. I argue that according to these approaches, there might arise some difficulties concerning the causal function of conscious intentions in individual and joint action, postulated in Pacherie’s social creation myth. Discussing a potential solution, how to understand the “causal” role of conscious intentions in the social creation myth despite those limitations, this commentary could serve as a complementary approach to the social creation myth of Pacherie. I want to argue that a creation myth cannot answer the relevant question, *how* conscious intentions play a role in our acting, without considering the nature of conscious intentions and thereby simultaneously focusing on their causal role.

2 Different myths about conscious intentions

According to [Bratman](#)’s pragmatic teleological creation myth (1987), intentions are future-directed action plans, offering humans the capacity to “become temporally extended agents” ([Pacherie this collection](#), p. 3). By forming an intention, which is inert and stable, we are able to predict the future and our future planning and form the basis for further intentions. Pacherie criticizes the future-directedness of conscious intentions and says that the pragmatic account of Bratman is incomplete, as it leaves non-pragmatic and present-directed intentions out of sight. Answering to the non-pragmatic function of conscious intentions, [Anscombe](#)’s teleological creation myth is (1963). [Anscombe](#) (1963) gives the whole debate about conscious intentions a highly interesting epistemic turn; her idea of conscious intentions is that they “provide us with a special kind of self-

knowledge” ([Pacherie this collection](#), p. 5). Her view of conscious intentions is that they provide immediate knowledge of our intentional actions as they provide an immediate non-observational and direct access to the content of our intention. [Velleman](#)’s myth about the function of conscious intentions is different (2007). He proposes that conscious intentions are a spandrel and do not have a teleological function on their own, they are a mere “accidental by-product” ([Pacherie this collection](#), p. 6) of two features of human nature: curiosity and self-awareness. From these features arises the concept of intentions that allows human individuals to understand their actions in the world. Pacherie argues that Velleman’s approach only shifts the problem of the function of conscious intentions to the function of curiosity and self-awareness.

Pacherie’s suggestion is an approach based on empirical knowledge and conceptual considerations about motor cognition. The central element is the suggestion that conscious intentions have a function in motor control. She proposes a three-step hierarchical concept of generation and control of motor actions, developed elsewhere ([Pacherie 2008](#)). Motor actions are controlled in an inverse and forward model, comparing error signals on different levels with each other. On the highest level I-Intentions are formed, referring to an abstract goal. These I-intentions allow for the selection of a fitting motor programme, the P-intention. Based on the P-intention the action underlies an online motor control via the M-Intention.

Although providing evidence for unconscious motor control on the lowest level, [Pacherie](#) argues that a control function of intentions cannot be denied and remains a “central function of intentions” ([this collection](#), p. 9), mainly on the highest level. Unconscious corrections are sufficient for small misalignments on the lowest level, whereas conscious intentions are necessary in the case of large misalignment between the different levels of motor control. [Pacherie](#) declares that “[i]n such case[es] of large

misalignment, error signals would propagate upwards, we would become aware of them, and would shift to a conscious, intentional compensation strategy” ([this collection](#), p. 10). Pacherie also offers a new definition of intentions. She thinks

of monitoring and control processes as intrinsic to intentions, that is, of intentions as encompassing not just representations of goals but also as a specific set of monitoring and control processes, organizing and structuring motor processes that themselves generate movements. ([Pacherie this collection](#), p. 10)

To summarize, one can understand Pacherie’s conscious intentions as having a causal function.

One step further Pacherie suggests that conscious intentions have a coordinative and communicative function in joint action on the basis of her idea that they arise through a hierarchical action control mechanism. Joint action between humans needs a common goal, and success of the joint action is based on our capacity to coordinate actions and share goals, and also to correct and control the individual actions according to the co-agent’s actions. Shared actions can, in analogy to the hierarchical model of individual motor control, be controlled on a sub-conscious low-level. In planned action, however, a hierarchical high level of motor control is needed with which agents represent other’s actions and control their own actions according to a shared goal. Mechanisms for joint action discussed in recent empirical science focus on a perceptual framework with joint attention and allocentric spatial orientation ([Tomasello & Carpenter 2007](#); [Böckler et al. 2011](#) cited from [Pacherie this collection](#)). The question however, is whether this perceptual information is sufficient for successful joint action. Pacherie concludes that *the conscious intention* is necessary to control intra-individual and inter-individual alignment of actions. One major aspect in joint action is communication of joint goals—so the *conscious* intentions help us to communicate our intentions to others and the other way round, to receive information about the inten-

tion of others and to represent them. The influence of other’s intentions then guides our own intentions and the following actions.

After this overview over the different creation myths, we should think about the concept of a “myth” itself. A myth in general tries to find an explanation for a phenomenon that we cannot entirely understand. There seems to be a missing piece of knowledge, a gap, which is filled with an idea—the myth. Defining characteristics of a myth since ancient philosophy are its narrative or descriptive character, without being completely irrational. A myth in Plato’s sense can neither be falsified nor empirically verified ([Partenie 2014](#)). So a myth offers a possible explanation about a phenomenon, without making a claim about truth and without offering a potential empirical approach to the content of the myth. A creation myth about specific functions of conscious intentions is developed, as they seem so unquestionable in our everyday life, and nevertheless, we do not understand, why we have them. The myth—however—does not necessarily need to fit the rules of the physical world.

The social creation myth endows conscious intentions with the important function of a structuring and organizing part in motor action. To my mind, Pacherie develops even more than a myth. The above mentioned characteristics of a myth do not fit Pacherie’s empirically based approach. She wants to understand and explain the function of conscious intentions, and her myth wants to *prepare* us for such a deeper understanding. That is an important step, yet it brings certain difficulties. An explanation has to fit into the framework of current scientific knowledge. Most creation myths and most explanation myths make some implicit or explicit assumptions about the nature of conscious intentions, so do the above-described myths. To make full use of Pacherie’s contribution, we now should begin by *adding* constraints. [Pacherie](#) herself knows these limitations and discusses some of them in her recent paper ([2014](#)). What I want to add is a step-by-step-comparison of the empirical and analytical, metaphysical constraints and her hierarchical model in the following sections.

3 Conceptual constraints: The problem of mental causation

Folk psychology tells us that our bodily movements, our actions, are guided by our intentions. One prominent conception of this assumption was developed as part of a non-reductive approach taken by Searle. For Searle, an action is “a causal and intentional transaction between mind and the world” (1983, p. 88). Searle distinguishes between two kinds of intentions, a *prior intention* and an *intention-in-action*. This distinction serves to preserve the difference between an intention as a basic idea or plan, preceding an action and an intention while carrying out an action. If a person *P* has a prior conscious intention for an action *A*, *P* has a representation of *A* without actually doing *A*. This is—according to Searle—a deliberative state and represents the *action as a whole*. Contrary, the *intention-in-action* occurs simultaneously with the action, representing the actual *conditions of the action*. Conditions can be regarded as certain steps, an action needs to be carried out. *P* has an intention-in-action-while *A*. But, the prior intentions are causally responsible for the intention-in-action and the action itself (Searle 1983).

This is, what—to my understanding—Pacherie’s social creation myth stresses as well. At the beginning of Pacherie’s paper about conscious intentions, a crucial point is made about the causal connection between intentions and actions: “Roughly, the notion on intentions is of a mental state that represents a goal (and a means to that goal) and contributes, through the guidance and control of behaviour, to the realization of what it represents” (this collection, p. 1). Her considerations about intentionality are about practical intentionality, as they concern conscious intentions in action, not only theoretical or cognitive intentions as mental representations. On the level of metaphysics, her statement could be interpreted along the lines of two kinds of property dualism. First it could be interpreted in a functionalist way in which conscious intentions are *abstract* mental properties possessing a causal role for our actions, in which they have a neuronal realisation

or implementation in the background (Lycan 1987; Clark & Chalmers 2002). Secondly, it could be interpreted in a way that declares conscious intentions to be non-reducible, non-physical mental properties, to be local instantiations, which are preceding or accompanying our actions. This notion of conscious intentions describes the conscious intention as a supervening or emerging mental property, which has a physical basis but is not identical or causally dependent with it (Davidson 1980; Kim 1998).

These non-reductional understandings are challenged from a variety of directions. Psychophysical correlations can also be conceptually interpreted using metaphysical models like identity theory (Feigl 1967; Place 1960; Smart 1959) or reductive or eliminative materialism (Churchland 1981), leaving no room for any causal function of conscious intentions. So, according to the most popular models developed after World War II, no conscious intention is a distinct mental entity or an ontological substance in a Cartesian sense. Nevertheless, different assumptions about the nature and the causal function of conscious intentions do exist. To present these in a provocative and simplified way, conscious intentions can either be a mental phenomenon in a physical world and have a causal role (compare: functionalism or non-reductive approaches), or they are causally irrelevant, since they are a by-product of our actions, an epiphenomenon, and as such non-existent (compare this to eliminative materialism).

I now want to focus on non-reductive approaches, as they seem to be relevant for the understanding of Pacherie’s social creation myth. Non-reductive approaches, which are supported by the common sense of conscious intentions and intentional action, and which all suggest that our conscious intention initializes the following action, however, might lead to a dilemma. As Heil and Mele put it: “We confront a dilemma. Either we concede that ‘purposive’ reason-giving explanations of behavior have only a pragmatic standing, or we abandon our conception of the physical domain as causally autonomous” (Heil & Mele 1995, v). The intuition that mental states have causal power is opposed by the rule of causal closure of the

physical world. Kim develops one notion of causal closure with the argument of causal exclusion and supervenience in his essay: *Mind in a Physical World* (1998). In a physical world, in which we do not have a complete physical monism but a non-reductive physicalism with supervenience, two premises are true: (1) every mental property M needs a physical basis P, which is sufficient for the existence of M and on which it supervenes and (2) every physical effect has a sufficient physical cause. Suppose M causes another mental property M*. M* has a physically sufficient basis P*. The problem which arises then is that M and P* as a causally sufficient basis are both responsible for the occurrence of M*, so M has to cause the physical basis P* of M* in a way of mental-to-physical causation. This result conflicts with the premise of causal closure of the physical world, according to which every physical event that has a cause has a *physical* cause (P causes P*). Facing now an over-determination of P*, with two different causally sufficient events *competing* for the causation of P* (M and P), and as P is causally sufficient for M, P seems to be causally sufficient for P*, and M does not have any causal effect itself. A mental phenomenon, according to this view, seems to be causally irrelevant. This is a rather short version of the causal closure-argument; the whole discussion about mental causation and causal closure cannot be displayed here (for an overview see e.g., Heil & Mele 1995). But the causal closure argument seems to be a problem for both: non-reductionist and functional approaches.

What consequences have to be drawn from these considerations about the causal function of conscious intention? Asking for the function of conscious intentions, the different creation myths face the problem of causality in a different way. Both the pragmatic (Bratman) and the epistemic (Anscombe) creation myths are set on a rather abstract mental level of description. Now, coming back to the two-level distinction of intentions introduced by Searle, both teleological myths are about prior intentions. One could say that neither teleological myths require any assumptions about causality, as they do not involve a mind-world directed causality, but

rather an intra-mental causality. In the pragmatic creation myth, intentions are preceded and followed by other intentions or intentions to act. Intentions are merely theoretical intentions as they only have a representational character. We can think of the pragmatic creation myth without any real action going on, as an abstract framework for an explanation of the existence of conscious intentions. The epistemic creation myth does not affect the debate about mental causation either. As only a correlation between conscious intention and action is necessary for the epistemic creation myth, it only draws conclusions about self-awareness and does not make any claim about a causal relation of this self-awareness and an action. In Vellemann's view, conscious intentions are a spandrel, a by-product. This model does not imply any explicit claim about causality either. One could go even one step further and postulate that these myths only address the structure of phenomenological experience of conscious intentions and not the intention itself.

In Pacherie's social creation myth, one cannot deny a causal role of conscious intentions any more. This is what I outlined above, referring to Pacherie's definition about conscious intentions. Intentions are not "just" a representation of abstract goals, but of ongoing control and they structure motor processes and "themselves generate movements" (Pacherie this collection, p. 10). Even more, if conscious intentions are needed to modify a joint action, the perception of goal-directed movements of others leads to a mental representation of this action and the formation of a conscious intention for another action follows from this. The problem with conscious intentions in Pacherie's social creation myth could arise when we understand—as outlined above—the I-Intention as purely functional or supervening mental properties in a non-reductive metaphysical framework. Regarding Searle's distinction between *prior intentions* and *intentions-in-action*, I assume that in Pacherie's model the I-intentions could be regarded as *prior intentions* and the P- and M-intentions rather correspond to *intentions-in-action*. To be more precise in this comparison we should talk about the experience of an inten-

tion as conscious mental representation (I-intention). As outlined above, one major analytical constraint against this understanding is the argument of causal closure. If the conscious intention (the I-intention) as a mental phenomenon or a mental representation has a causal function in action, we have to accept downward causation to understand this (which would be against the rule of causal closure). If the I-intentions only supervenes or emerges from its neuronal activity, or is identical with it, then the intention as a conscious mental representation is causally irrelevant and not necessary for the function of motor control. My claim here is that Pacherie's social creation myth needs the causal function of conscious intentions as mental representations to work. Yet requires that we accept the idea of mental causation. As long as the social creation myth is only a myth, we can break the rules of causal closure easily and just offer the gist or general structure of a potential explanation about the function of conscious intentions. Yet if the myth is an explanation, it has to fit the rules of causal closure, and we have to reconsider either the myth or our understanding of causal closure. Last, we could try to create a myth fitting our physical knowledge, yet have to deny the causal effect of the conscious intentions in motor control.

4 Empirical constraints: Current neuroscientific knowledge about the status of conscious intentions

The question about the function of conscious intentions cannot be answered by conceptual considerations alone. The status of practical conscious intentions can be analysed in motor action—as it is done by Pacherie as well—but not only on the level of theoretical hierarchical models of motor initiation and control but on a mere neurophysiological level. Let's begin with a classical example—the Libet-experiments (Libet et al. 1983, 1985) and their modified versions by Haggard & Eimer (1999). Libet and his colleagues designed an experiment to investigate the temporal connection between a voluntary motor activity and the conscious decision—the conscious intention—for this action. They in-

structed their test persons to voluntarily move their hand and to detect the time at which the urge or the conscious intention to move their hand developed. In parallel, muscle activity was detected via electromyography (EMG) and the readiness potential, a neuronal potential at the beginning of a motor action, was recorded using electroencephalography (EEG). Libet and his colleagues found that the readiness potential can be detected in average 350ms earlier than the test persons experienced the *urge to move* and postulated that according to this finding the decision to move cannot be causally responsive for the action due to a time-based difference. One interpretation of the experiments is that neuronal activity (the readiness potential for the motor activity) occurs before the conscious knowledge of the action itself. So, the conscious intention itself cannot be responsible for a volitional motor action as it occurs later than the subconscious neuronal changes. These findings initiated an on-going debate about the connection between motor activity and the being conscious about this activity, with many neuroscientists supporting the initial hypothesis. Haggard & Eimer detected a lateralized readiness potential (1999). Libet's experiment has been replicated in various alternations, supporting the view that conscious intentions follows pre-conscious brain activity fitting to the movement (Trevena & Miller 2002; Siguru et al. 2004; Rigoni et al. 2011). Similar results were shown for the inhibition of an action (Filevich et al. 2013). fMRI studies (Lau et al. 2004; Soon et al. 2008; Haggard 2008) and transcranial magnetic stimulation-studies postulated a neuronal preceding to motor action similar experimental paradigm (for reviews see Haggard 2005; Shields 2014). One recent fMRI-study for example, reported successful prediction of free choices (addition or subtraction) in the study persons due to fMRI data analysis (Soon et al. 2013). Even single-cell recording in humans—as an objective approach to the self-initiated action—detected neuronal recruitment prior to the intention to act (Fried et al. 2011). The conclusion of above-mentioned experiments frequently is, that the conscious intention of a movement is either an illusion or a post-hoc attribution, generated by the movement itself.

On a conceptual level, there exist other models about conscious motor control besides Pacherie's hierarchical model. An important idea is the idea of intentional binding (Haggard et al. 2002), where an intentional action is causally linked with a certain sensory outcome. In this case, the action and its subsequent effect are perceived as being closer together in time, this generates the phenomenology of causing and independently originating the action, without an actual causal function of the conscious intention. Another current neurobiological theory of motor control is often referred to as comparator model (Frith et al. 2000). Every action consists of two kinds of representations: inverse models that specify motor commands according to sensory perception and forward models that represent the predicted sensory consequences of the movement. When a comparator signals that the sensory consequences of the movement match those predicted by the forward model, we experience this action as consciously intended. Here again, the conscious intention is not causally responsible for the action.

Transferred to the terminology of intentions, this interpretation could mean that a *prior intention* (or I-intention) cannot be causally responsible for an *intention-in-action* (lower level intention) as the neuronal activation pattern for the prior intention was earlier detected than the intention was reported as conscious. What would be the conclusion regarding the social creation myth? As a conscious intention itself—according to the above mentioned interpretation—is not regarded to be causally responsible for the initiation of a motor activity (only the subconscious neuronal activity is responsible) the conscious mental representation of a motor activity in individual or joint action is not causally involved in the processes of motor control. The function of conscious intentions in the social creation myth either stays a myth, as it contradicts the empirical findings, or the myth fits the nature of conscious intentions and we have to reconsider the interpretation of the experiments.

To support the later alternative, one recent study using transcranial magnetic stimula-

tion, a method which allows generating movements by transcranial stimulation of the neurons of the motor cortex, postulated that motor activity is initiated by conscious intentions. A transcranial stimulus was set in the right motor cortex and introduced a tiny muscle twitch, only recordable by EMG. When test persons intended to move their left hand prior to the transcranial stimulus, the transcranial-induced involuntary movement induced a stronger visible motor response. The authors postulated that the conscious intention prepares volitional motor actions by increasing the excitability of the cells in the motor cortex that can produce the movement intended (Zschorlich & Köhling 2013).

There are further some major limitations to the studies, e.g., the subjectivity of the report of the urge to move, and the highly artificial/constructed experimental situation in which the intentional action is carried out. One common objection against an interpretation of the data in the way of Libet and colleagues is that conscious intentions (e.g., the *prior intentions*) are not comparable to the urge to move in an experimental setting but rather are comparable to the decision to participate in the whole experiment. The urge to move would rather be an *intention-in-action* and by this not comparable to a conscious deliberation about an action. Following from the data, a conscious intention is unnecessary or irrelevant (as it occurs “too late”) in conscious motor initiation and control could be a too far-reaching conclusion.

5 The problem of causality and the search for a new myth

The aims of the commentary were first to understand, why Pacherie's social creation myth is more than a myth. Second, I elucidated whether it could, in principle, lay the foundations for an explanation based on and in line with philosophical and experimental ideas about mental causation. This discussion was based on the more general question: do conscious intentions have a causal function in the world? To my mind this question cannot yet be answered conclusively, at least according to our current

knowledge. Postulating a lack of causal function of conscious intentions, as based on analytical considerations and empirical data, might be the only possible solution of the problem. The argument from causal closure postulates that a conscious intention as a mental phenomenon is causally irrelevant, because it is not needed to explain a following physical phenomenon. The experimental data might suggest that an intention becomes conscious only after the neuronal activity is detected. Yet, there still is the strong experience of a causal function for our behaviour.

Now, I want to summarize the problems for the social creation myth, based on the above mentioned discussion and I want to consider possible ways to keep and develop the social creation myth as a potential explanation about the function of conscious intentions. The general question about the function of experienced conscious intentions, as [Pacherie](#) puts it, is the question about “the normative sense, in which having these functions confers benefits on intention-forming creatures that explains why these creatures have this capacity” ([this collection](#), p. 1). This general question is one of the interpretation and explanation of human nature and not a question about causality. The creation myths of Bratman and Anscombe mainly address the question of why we experience our intentions as conscious and goal directed. The question about real-world, physical causality seems unessential for a pragmatic or epistemic benefit for our being and self-awareness, because the pragmatic or epistemic benefit of conscious intentions arises from the experience of a conscious intention and not from its causal effect. The intentions remain theoretical intentions or mental representations and no downward causality is needed. This does not mean that they cannot have a specific and more complex function, but a strong claim about a localized control-function in motor action is simply not possible. In addition, the epistemic and the pragmatic creation myth as well as conscious intentions considered as a spandrel remain “narrative” accounts and even if they would break the causal closure of the physical world, this would not matter in the context of a myth.

Pacherie’s social creation myth first seems to be of a similar kind, explaining human nature and human interaction on the basis of mutual representation of others’ actions and formation of joint actions, which do not necessarily have to be causal for joint action, but only for communication intentions and our understanding joint action. The social creation myth is based on the conceptual, hierarchical model of motor initiation and control. It explains conscious intentions not only in a teleological way, but in an analytical way. It is about practical intentionality. Yet, this confronts it with neuroscientific findings and philosophical considerations about causality:

- Conscious intentions in Pacherie’s social creation myth exert an organizing and structuring function in the motor process and therefore might have a causal function.
- According to standard metaphysical models for psychophysical relations, the conscious intentions in the myth could be interpreted as a non-reducible mental phenomenon. But if this is the right interpretation, we are confronted with the argument of causal closure and they are either causally irrelevant or we have to deny causal closure of the world.
- According to neuroscientific data, we only know little about the nature of conscious intentions, yet nevertheless we have a strong general trend underlying empirical research, a trend that increasingly supports the assumption of a generation of the wanting or the urge to move from neuronal activation, simultaneously or after, but not prior to the movement.

What does this mean for the social creation myth? Regarding the outlined considerations about causality, the problem of the social creation myth about the function of conscious intentions can be solved in different ways. Either we could regard it as a myth in line with the teleofunctional creation myths, only trying to answer the “why”-question about conscious intentions and leaving questions about causality aside. This could sidestep the problem of causality in an easy yet unsatisfy-

ing way. But if we stick to a myth without acknowledging the physical rules of the world we live in, then we will never achieve more detailed knowledge about the nature and the function of conscious intentions. There will be no epistemic progress after the formulation of the myth itself.

Or we try to preserve Pacherie's approach and keep searching for an explanation about the function of conscious intentions. Yet, if conscious intentions have a structuring and organizing function in individual and joint motor action but—according to the common interpretation of above mentioned empirical data—cannot have distinct causal function, how else can the function be described?

One possible solution is, that we might have to overcome the problem of causality in another way. Most interpretations of neuroscientific experiments and the analytical argumentation of causal closure are based on a temporal, linear one-way causality in the way that A causes B because A precedes B. Additionally, one single intention is typically regarded as the cause of the action in a quasi-linear model. My claim is that the common interpretation that a conscious intention—qua being conscious—can only be causally relevant if the conscious intention precedes the motor action, has to be revised. A first motivation for this claim is the fact that there are multiple theoretical and practical limitations regarding the experiments themselves (e.g., [Mele 2011](#); [Radder & Meynen 2012](#); [Pacherie 2014](#)).

But even if the common conceptual interpretation was right, there might be a further terminological problem. In the whole debate about conscious intentions in the social creation myth, we seem to assume that there must be a certain effect of the *being conscious* of the intention. *Because* an intention is conscious, it has an effect to align and control motor action. If it was not conscious, it would not have this effect. To overcome these problems in the debate of the function of conscious intentions, I suggest that a different concept of causation should be considered. This alternative refers to a parallel generation of a con-

scious intention and movement planning. As it is a parallel process and we might be confronted with two aspects of one and the same process, the conscious intention neither precedes nor follows the action generation, but occurs simultaneously and both are influenced reciprocally ([Desmurget 2013](#)). Even further steps may have to be taken. It has been postulated that we cannot trace back the motor action onto one I-Intention in a linear model or to one single place of neuronal activity in the brain. We rather face a semi-hierarchical, parallel and dynamic network from which the motor action arises, without single, identifiable conscious intentions in a direct line of causality but rather fluctuating activity ([Schurger 2014](#)). This would mean that various intentions exist and each of them can influence, control and generate motor action on a neuronal level in parallel, these intentions are among others generated through the observation and interaction with others. Multiple goal representations might form a context for each other. On a conceptual level there would be different I-intentions and different motor programmes going on at the same time. But let us assume that only some of these I-intentions are conscious. Being conscious, for Pacherie, is a necessary condition to exert a motor function and to align actions with others; being conscious is necessary for the causal role in her creation myth. Maybe the function of being conscious could exert a certain weight to an I-intention, not in the way of a linear causality but in a way of dynamic modelling a given social context.

This could save the social creation myth and sheds new light on the interpretation of neuroscientific findings. Whether or not this move answers the question about the function of conscious intentions remains open. The aim should be to further integrate the analytical definitions of mental phenomena and mental causation into neuroscientific research about conscious intentions and try to find a working definition and a concept of what a conscious intention is like. The focus should be on the function of practical conscious intentions and analyse their causal role and function for the hu-

man nature on a neuronal level. Maybe future attempts to arrive at a satisfactory explanation should try to address the causal power of a conscious intention *while* being conscious and not *because of* being conscious.

6 Conclusion

So, do we need a creation myth after all? One thing is certain: conscious intentions unquestionably exist in our experience. We have at least the phenomenal experience of a conscious intention in our acting. As conscious intentions seem so relevant for our human nature we do need a myth about them. But we need even more. Pacherie's social creation myth to my mind is more than a myth; it is one approach, which combines empirical knowledge with a myth about the function and its history. I have only analysed the question of causality from an empirical and metaphysical point of view and its relevance for the social creation myth. In conclusion, we might have to satisfy some further analytical and empirical constraints. Yet, just denying any function of the experience of conscious intentions due to some experimental data or analytical considerations seems premature. A possible solution could be the reconsideration of the concept of causality, to find an explanation of the function of conscious intentions in individual and joint action. Maybe the creation myth and the experimental approach have to be adjusted and be brought together in concept and content, in order to understand the deeper function of conscious intentions. The search for a creation myth should start with creation facts. These facts should help us to elucidate why and how intentions are conscious or at least achieve their phenomenal character, to define the neural correlates or neural correlation in terms of self-organizing, dynamic networks underlying conscious intentions and the causal function in human action, without the limitations of temporal or linear causality and in a more realistic framework of intentional action.

References

- Anderson, M. L. (2015). Beyond componential constitution in the brain: Starburst amacrine cells and enabling constraints. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Anscombe, G. E. M. (1963). *Intention*. Oxford, UK: Blackwell.
- Bratman, M. (1987). *Intention, plan, and practical reason*. Cambridge, MA: Harvard University Press.
- Böckler, A., Knoblich, G. & Sebanz, N. (2011). Giving a helping hand: Effects of joint attention on mental rotation of body parts. *Experimental Brain Research*, 211 (3-4), 531-545. [10.1007/s00221-011-2625-z](https://doi.org/10.1007/s00221-011-2625-z)
- Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy*, 78 (2), 67-90.
- Clark, A. & Chalmers, D. (2002). The extended mind. In D. Chalmers (Ed.) *Philosophy of mind* (pp. 643-651). New York, NY: Oxford University Press.
- Craver, C. F. (2015). Levels. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Davidson, D. (1980). Mental events. In L. Foster & J. W. Swanson (Eds.) *Experience and theory* (pp. 79-101). Amherst, MA: University of Massachusetts Press.
- Desmurget, M. (2013). Searching for the neural correlates of conscious intention. *Journal of Cognitive Neuroscience*, 25 (6), 830-833. [10.1162/jocn_a_00368](https://doi.org/10.1162/jocn_a_00368)
- Feigl, H. (1967). *The "mental" and the "physical": The essay and a postscript*. Minneapolis, MN: University of Minnesota Press.
- Filevich, E., Kühn, S. & Haggard, P. (2013). There is no free won't: Antecedent brain activity predicts decisions to inhibit. *PLoS One*, 8 (2), e53053. [10.1371/journal.pone.0053053](https://doi.org/10.1371/journal.pone.0053053)
- Fried, I., Mukamel, R. & Kreiman, G. (2011). Internally generated preactivation of single neurons in human medial frontal cortex predicts volition. *Neuron*, 69, 548-562. [10.1016/j.neuron.2010.11.045](https://doi.org/10.1016/j.neuron.2010.11.045)
- Frith, C. D., Blakemore, S. J. & Wolpert, D. M. (2000). Abnormalities in the awareness and control of action. *Philosophical Transactions of the Royal Society of London B*, 355 (1404), 1771-1788. [10.1098/rstb.2000.0734](https://doi.org/10.1098/rstb.2000.0734)
- Gallagher, S. (2005). *How the body shapes the mind*. Oxford, UK: Oxford University Press/Clarendon Press.
- Haggard, P. (2005). Conscious intention and motor cognition. *Trends in Cognitive Sciences*, 9 (6), 290-295. [10.1016/j.tics.2005.04.012](https://doi.org/10.1016/j.tics.2005.04.012)

- (2008). Human volition: Towards a neuroscience of will. *Nature Reviews Neuroscience*, 9, 934-946. [10.1038/nrn2497](https://doi.org/10.1038/nrn2497)
- Haggard, P. & Eimer, M. (1999). On the relation between brain potentials and the awareness of voluntary movements. *Experimental Brain Research*, 126 (1), 128-133.
- Haggard, P., Clark, S. & Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature Neuroscience*, 5, 382-385. [10.1038/nn827](https://doi.org/10.1038/nn827)
- Heil, J. & Mele, A. (1995). *Mental causation*. Oxford, UK: Oxford University Press.
- Kim, J. (1998). *Mind in a physical world. An essay on the mind-body problem and mental causation*. Cambridge, MA: MIT Press.
- Lau, H. C., Haggard, P. & Passingham, R. E. (2004). Attention to intention. *Science*, 303 (5661), 1208-1210.
- Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences*, 8, 529-566. [10.1007/s00221-011-2625-z](https://doi.org/10.1007/s00221-011-2625-z)
- Libet, B., Gleason, C. A., Wright, E. W. & Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). *The unconscious initiation of a freely voluntary act*. *Brain*, 106, 623-642.
- Lycan, W. (1987). *Consciousness*. Cambridge, MA: MIT Press.
- Mele, A. (2011). Libet on free will: Readiness potentials, decisions, and awareness. In W. Sinnott-Armstrong & L. Nadel (Eds.) *Conscious will and responsibility* (pp. 23-33). Oxford, UK: Oxford University Press.
- Pacherie, E. (2008). The phenomenology of action: A conceptual framework. *Cognition*, 107 (1), 179-217. [10.1016/j.cognition.2007.09.003](https://doi.org/10.1016/j.cognition.2007.09.003)
- (2014). Can conscious agency be saved? *Topoi*, 33 (1), 33-45. [10.1007/s11245-013-9187-6](https://doi.org/10.1007/s11245-013-9187-6)
- (2015). Conscious intentions: The social creation myth. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Partenie, C. (2014). Plato's myths. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. <http://plato.stanford.edu/archives/sum2014/entries/plato-myths>.
- Place, U. T. (1960). Materialism as a scientific hypothesis. *Philosophical Review*, 69 (1), 101-104.
- Radder, H. & Meynen, G. (2012). Does the brain "initiate" freely willed processes? A philosophy of science critique of Libet-type experiments and their interpretation. *Theory & Psychology*, 23 (1), 1-19. [10.1177/0959354312460926](https://doi.org/10.1177/0959354312460926)
- Rigoni, D., Kühne, S., Sartori, G. & Brass, M. (2011). Inducing disbelief in free will alters brain correlates of preconscious motor preparation: The brain minds whether we believe in free will or not. *Psychological Science*, 22 (5), 613-618. [10.1177/0956797611405680](https://doi.org/10.1177/0956797611405680)
- Schurger, A. (2014). Intentions and voluntary actions: Reframing the problem. *Cognitive Neuroscience*, 5 (3-4), 213-214. [10.1080/17588928.2014.950214](https://doi.org/10.1080/17588928.2014.950214)
- Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge, UK: Cambridge University Press.
- Shields, G. R. (2014). Neuroscience and conscious causation: Has neuroscience shown that we cannot control our own actions? *Review of Philosophy and Psychology*, 5 (4), 565-582. [10.1007/s13164-014-0200-9](https://doi.org/10.1007/s13164-014-0200-9)
- Siguru, A., Daprati, E., Ciancia, S., Giroux, P., Nighoghossian, N., Posada, A. & Haggard, P. (2004). Altered awareness of voluntary action after damage to the parietal cortex. *Nature Neuroscience*, 7 (1), 80-4. [10.1038/nn1160](https://doi.org/10.1038/nn1160)
- Smart, J. J. C. (1959). Sensations and brain processes. *Philosophical Review*, 68, 141-156.
- Soon, C. S., Brass, M., Heinze, H.-J. & Haynes, J.-D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, 11, 543-545. [10.1038/nn.2112](https://doi.org/10.1038/nn.2112)
- Soon, C. S., Hanxi He, A., Bode, S. & Haynes, J.-D. (2013). Predicting free choices for abstract intentions. *Proceedings of the National Academy of the USA*, 110 (15), 6217-6222. [10.1073/pnas.1212218110](https://doi.org/10.1073/pnas.1212218110)
- Tomasello, M. & Carpenter, M. (2007). Shared intentionality. *Developmental Science*, 10 (1), 121-125. [10.1111/j.1467-7687.2007.00573.x](https://doi.org/10.1111/j.1467-7687.2007.00573.x)
- Trevena, M. & Miller, J. (2002). Cortical movement preparation before and after a conscious decision to move. *Conscious and Cognition*, 11 (2), 162-90. [10.1006/ccog.2002.0567](https://doi.org/10.1006/ccog.2002.0567)
- Velleman, D. (2007). What good is a will? In A. Leist & H. Baumann (Eds.) *Action in context* (pp. 193-215). Berlin, GER: de Gruyter.
- Zschorlich, V. R. & Köhling, R. (2013). How thoughts give rise to action: Conscious motor intention increases the excitability of target-specific motor circuits. *PLoS One*, 8 (12), e83845. [10.1371/journal.pone.0083845](https://doi.org/10.1371/journal.pone.0083845)

The Causal Role(s) of Intentions

A Reply to Andrea R. Dreßing

Elisabeth Pacherie

In her commentary ([Dreßing this collection](#)) on my target article ([Pacherie this collection](#)), Dreßing suggests that the story I offer is not just a creation myth but also an attempt to give an explanation of the function of conscious intentions in the physical world and as such answerable to both metaphysical and empirical constraints. Here, I try to clarify which of my claims should be understood as simply speculations about the origins of our capacity of intentions and which I take to be empirical claims. In response to the metaphysical and empirical challenge Dreßing raises, I argue that Dretske's distinction between structuring and triggering causes may help us see how explanations in terms of physical properties and explanations in terms of mental properties may not compete but rather complement each other. I argue that this distinction may also help us assuage certain worries raised by neuroscientific findings.

Keywords

Causal exclusion | Conscious agents | Conscious intention | Creation myth | Intentional action | Intentions | Joint action | Mental causation | Neuronal correlates of intentions | Structuring causes | Triggering causes

Author

[Elisabeth Pacherie](#)

elisabeth.pacherie@ens.fr
Ecole Normale Supérieure
Paris, France

Commentator

[Andrea R. Dreßing](#)

andrea.dressing@uniklinik-freiburg.de

Klinik für Neurologie und
Neurophysiologie
Universitätsklinikum Freiburg
Freiburg, Germany

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Introduction

In her commentary, [Andrea Dreßing](#) ([this collection](#)) suggests that I might have been too timid in calling the story I tell about the social function of intentions in my target article ([Pacherie this collection](#)) a creation myth. She encourages me take a bolder stance, claiming that the story I offer is not just a myth but also an attempt to give an explanation of the function of conscious intentions in the physical world. Indeed, part of my story is intended as more than a myth and so my first task here will be to clarify where I

draw the line between empirical claims and myths.

Dreßing also points out that an explanation, as opposed to a mere myth, has to fit into the framework of current scientific knowledge and is therefore subject to both metaphysical constraints and empirical constraints. I concur. In what follows, I will argue, however, that my general predicament with regard to conceptual or metaphysical constraints is not so different from the predicament of the other myth-tellers I

discuss in my article, as Dreßing suggests. Nor indeed is it direr than the predicament all philosophers of mind working within a naturalistic framework face. Finally, certain empirical findings have been interpreted as showing that conscious intentions play no role in action initiation. I also try to address this challenge.

2 Myths vs. empirical claims

In my target article, I use the phrase “creation myth” first as a dramatization device. Typically, we do not feel the urge to formulate myths about things we deem insignificant. Talking of a social creation myth was thus a way of emphasizing the importance of the social function of intentions, a function largely neglected in traditional accounts of intentions. Second, I also wanted, following Velleman (2007), to convey a note of caution. A myth, as Dreßing points out, can neither be falsified nor empirically verified. It offers a possible explanation about a phenomenon, without making a claim about truth. But I perhaps wasn’t clear enough what I was trying to be cautious about and where I drew the line between empirical claims and ultimately unverifiable explanations. So let me now draw this line more firmly.

To do this, let me distinguish three different questions about intentions and examine how they may relate. The three questions are: *what* roles or functions (in a non-teleological sense) do intentions play in human agency? *How* can intentions play these roles? *Why* do we have intentions in the first place? In my view the *what*- and *how*-questions are both empirical questions for which mythical answers won’t do. The *why*-question, as I understand it, is a question about the origins of capacity for intention. How come we have such a capacity? Why was it established?

The focus of the account I proposed, as well as the focus of the alternative accounts by Bratman (1987), Anscombe (1963), and Velleman (2007) with which I contrast it in my article, is on the *what*- and *why*-questions. However, I offered my story as a creation myth only to the extent that it was meant to address the *why*-question. As answers offered to the *what*-

question, my claims were meant as empirical claims. I take it that the claims made by Bratman, Anscombe, and Velleman about the epistemic and pragmatic functions of intentions, when understood as answers to the *what*-question, should also be interpreted as empirical claims.

Now, how do the *what*- and the *why*-questions relate? One way to relate them is by assuming that intentions do not just have a function or functions in a value-neutral sense—things that they do—but a teleofunction in the evolutionary sense, that is, something that they do that confers some benefit or advantage on creatures with a capacity for intentions, and in this sense explains why these creatures have this capacity.

Velleman cautions us against this teleofunctional move. First, as his discussion of Bratman’s and Anscombe’s accounts makes clear, the *what*-question about intentions can be given complementary answers in terms of both pragmatic and epistemic roles, leaving us with several possible teleological stories. Second, Velleman also warns us against assuming direct links between answers to the *what*-question and answers to the *why*-question. The spandrel story he tells is meant to suggest that a capacity for intentions may only be a by-product of other capacities and thus that our capacity for intentions could be nothing more than an (admittedly very fortunate) accident. Finally, in calling his own story a creation myth as well, Velleman is also pointing out that our speculations about the origins of intentions are most likely beyond falsification or empirical verification.

Similarly, in offering my social function story as an answer to the *why*-question, I was not making a claim to truth. Rather, I was trying to broaden the terms of the debate to also include consideration of the social dimension of intentions. If we are considering what possible teleofunction intentions could have, then we should pay more attention to the benefits we derive from being able to act jointly in a flexible manner. If we are tempted by a story that views a capacity for intention as simply a by-product of more general capacities, then, among these more general capacities, we should pay

serious heed to our capacity for sociality and cooperativeness.

Turning now to the relations between the *what*-question and the *how*-question, I take it that the empirical standing of an answer to the *what*-question ultimately depends on whether this answer can be backed up by a convincing answer to the corresponding *how*-question. The validity of any empirical claim about the causal roles of intentions in human agency will remain in doubt unless one can see how it is at all possible for intentions to play these roles (Dreßing's metaphysical constraints), and it will also remain in doubt if appears to be in contradiction with well-established empirical facts (Dreßing's metaphysical constraints).

Since my claims about the functions of intentions qua answers to the *what*- rather than the *why*-question are intended as empirical claims, they are not insulated from these metaphysical and empirical worries. Let me address them in turn.

3 Metaphysical worries

Dreßing points out that my claim that intentions have a causal role to play in the online control of action confronts me with the problem of mental causation. She also suggests that this problem is more pressing for me than it is for the accounts of the functions of intentions proposed by Bratman and Anscombe. While I agree that the problem of mental causation is an issue for me, I disagree with her assessment that it isn't as serious a worry for these accounts.

First, let me clarify that when I talk about conscious intentions and their causal role, I am concerned with what [Ned Block \(1995\)](#) calls access consciousness rather than with phenomenal consciousness. In other words, my claims are about intentions qua conscious states exploiting and conveying information globally available in the cognitive system for the purposes of reasoning, speech, and high-level action control. My account thus faces the “easy” problems of consciousness rather than the “hard” problem ([Chalmers 1995](#)). I share [Chalmers](#) sanguine assessment

about phenomena pertaining to access consciousness:

There is no real issue about whether these phenomena can be explained scientifically. All of them are straightforwardly vulnerable to explanation in terms of computational or neural mechanisms. ([1995](#), p. 201)

This is not to say, however, that in confining oneself to phenomena of access consciousness one can eschew all metaphysical conundrums. In particular, as pointed out by Dreßing, the mere fact that Cartesian dualism has fallen out of favour and that the vast majority of philosophers and cognitive scientists are nowadays willing to embrace some form of materialist monism doesn't insure the dissolution of philosophical worries about mental causation. The version of the problem of mental causation that non-reductive physicalists, whatever their exact persuasion, are confronted with is the Causal Exclusion Problem: how could mental properties play a causal role given that they appear to be screened off by their physical realizers?

Dreßing argues that this problem is more pressing for my view than for the pragmatic (Bratman) or the epistemic (Anscombe) creation myths, the reason being that these latter two teleological myths are about prior intentions and that neither “require any assumption about causality, as they do not involve a mind-world directed causality, but rather an intra-mental mental causality” ([Dreßing this collection](#), p. 6). Dreßing also claims that Velleman's view does not imply any explicit claim about causality either, since on this view intentions are a spandrel or a by-product.

I disagree with this assessment for three reasons. First, as I explained in section 2, while the speculative character of these stories qua answers to the *why*-question may justify labelling them as creation myths, the stories also offer answers to the *what*-question. In that regard their claims about the epistemic or pragmatic roles of intentions should be taken as empirical claims. Thus, even if we go along with Velleman's claim that a capacity for intentions

is a spandrel and that the epistemic and pragmatic functions of intentions are not teleofunctions, they are nevertheless functions in the ordinary functionalist sense and we still need an explanation of how intentions can play these epistemic and pragmatic roles.

Second, the Causal Exclusion Problem is a problem for anyone espousing a non-reductive form of materialist monism,¹ whether their primary concern is with intra-mental causation or with mind–world causation. Suppose that a state *S* has the mental property *M* (e.g., the property of being an intention to go to London on Monday) and a physical basis *P*, suppose that *S'* has the mental property *M'* (e.g., the property of being an intention to buy a train ticket to London) and a physical basis *P'*, and suppose that *S''* has the mental property *M''* (e.g., the property of being a belief that one will go to London on Monday) and a physical basis *P''*. On a Bratmanian pragmatic account of intentions, I would want to be able to say that my intention to go to London on Monday causes, via further means-end reasoning, my intention to buy a ticket to London. But how can the mental property *M* of *S* play a causal role in bringing about a state *S'* with mental property *M'*, given that they appear to be screened off by the physical properties *P* and *P'*? Similarly, with regard to the epistemic function of intentions, how could I say that my intention to go to London on Monday causes my belief that I will go to London on Monday, given that mental properties *M* and *M''* appear to be screened off by the physical properties *P* and *P''*?

Third, while it is true that on Bratman's account future-directed intentions may only cause behaviour through the mediation of present-directed intentions, still Bratman insists that the whole point of having a capacity for intentions is to produce behaviour that contrib-

utes in the long run to our securing greater desire-satisfaction. Similarly, on a reliabilist reading of Anscombe's epistemic claim that intentions embody knowledge of our actions, they do so because intentions reliably cause what they represent. As Velleman puts it, “[u]nless an intention with the content ‘I’m going to move my toe’ reliably causes my toe to move, it won’t amount to practical knowledge” (Velleman 2007, p. 201). Thus, Bratman, Anscombe and Velleman cannot be exonerated from the task of explaining how mental states can cause behaviour.

With respect to the problem of mental causation, we are all in the same boat. The metaphysical standing of my account is no less or more precarious than the standing of these other accounts. Are we then all metaphysically doomed? Readers should not hold their breath; I have no new, unassailable solution to the problem of mental causation to offer. Yet, it would certainly be premature to claim that the problem of mental causation is insoluble. Many lines of response have been proposed and are currently being explored (for a review, see Robb & Heil 2014). I cannot discuss all these accounts here. Let me just say that the approach I find most congenial stems from Fred Dretske's work (Dretske 1988, 2004) on psychological explanations of behaviour. Dretske distinguishes between triggering causes and structuring causes, where a triggering cause is an event that initiates or triggers a causal chain of events, and a structuring cause the cause of the process or setup that makes a given triggering cause produce the effect it does. To take an example from Dretske (2004), moving a computer mouse is the triggering cause of cursor movement, but hardware and programming are the structuring causes of cursor movement. Dretske's central claim is that mental states and events are best analysed as structuring rather than triggering causes of behaviour. On this view there is no competition between physical and psychological or mental explanations, since they have different explananda. While the triggering physical properties explain bodily motion, i.e., explain why bodily motions occur at a certain point in time, the structuring mental properties explain

¹ While this issue is not at the heart of Bratman's preoccupations, I think we can safely assume that he would want his account of intentions to be compatible with physicalism. I won't dwell here on Anscombe's metaphysical view, except to say that she was no materialist herself but was also highly suspicious of Cartesian dualism (Anscombe 2008). Suffice it to say that many of the philosophers who nowadays embrace the view that intentions have an epistemic function, would want this claim to be compatible with a physicalist stance.

behavior, i.e., they explain why in circumstances of a certain sort, bodily motions of this kind rather than that kind are produced.

Much work remains to be done in order for us to understand more precisely how structural causes operate and in particular how they can do so in the dynamic way needed to account for the plasticity and flexibility of human behaviour. In this respect, Dretske's account remains largely under-developed (for recent work on this issue, see e.g., [Slors 2015](#); [Wu 2011](#)). Dretske's approach in terms of structuring causes has the great merit, however, of offering a potential solution to the Causal Exclusion Problem and to let us see how explanations in terms of physical properties and explanations in terms of mental properties may not compete but rather complement each other. As we will now see, thinking of intentions as structural causes of action rather than triggering causes can also help us assuage certain empirical worries.

4 Empirical worries

The claim that conscious intentions play a causal role in action production should be compatible with our best empirical knowledge on how action is produced. The main empirical worries this claim confronts come from neuroscientific findings that have been interpreted as showing that the time of onset of conscious intentions is not compatible with their being the initiators of actions.

The most famous of these experiments are Libet's studies on "readiness potential" ([Libet et al. 1983](#); [Libet 1985](#)). In these studies, subjects were asked to flex their wrist at will and to note when they felt the urge to move by observing the position of a dot on a special clock. While subjects were both acting and monitoring their urges (intentions, decisions) to act, Libet used an EEG to record the activity of prefrontal motor areas. On average, participants reported the conscious intention to act, which Libet called the W-judgement, about 200ms before the onset of muscle activity. By contrast, the EEG revealed that preparatory brain activity, termed by Libet type II readiness potential

(RP), preceded action onset by about 550ms. In other words, their brains started preparing the action at least 350ms before the participants became aware of their intention to act. This led Libet to the conclusion that the wrist-flexing actions in his experiments were not initiated by conscious intentions but were initiated instead by the (unconscious) RPs.

These experiments and Libet's interpretation of his findings have been widely discussed (see e.g., [Banks & Pockett 2007](#); [Bayne & Pacherie 2014](#); [Mele 2009](#); [Nahmias 2002](#); [Pacherie 2014](#); [Roskies 2011](#)) and commentators have pointed out a number of methodological problems with Libet's paradigm as well as conceptual problems with his interpretation of his results. Let me focus first on one methodological problem and one attempt to address it. I will then consider one conceptual problem

Libet argues that it is the RP rather than the conscious intention that initiate the agent's action. If RPs are the initiators of the action, there should be a robust correlation between them and the actions they cause: we should expect RP events to be "immediately" followed by the appropriate action, or, to put it the other way round, we should expect that when there is no movement, there is also no RP event. As several commentators have observed (e.g., [Mele 2009](#); [Roskies 2011](#)), the back-averaging techniques used in the experiment do not allow us to ascertain whether this is indeed the case. Because the RP on any one trial is obscured by neural noise, what is presented as "the RP data" is determined by averaging the data collected on a large number of trials. In order to compute this average, the EEG recordings on different trials need to be aligned, and this requires some fixed point that can be identified across trials. Since in Libet's experiments action onset serves as the needed fixed point for the alignment of EEG recordings, any RPs that are not followed by an action simply won't be measured, and so we don't know how robust the correlation between the RP and Libet-actions is.

In a recent experiment, Schurger and colleagues ([Schurger et al. 2012](#)) used a modified Libet task to circumvent the limitations of back-averaging techniques. Their aim was to

test the proposal that RPs correlate with pre-decision activity rather than, as Libet proposed, with activity that coincides with, or is subsequent to, the agent's decision. Schurger and colleagues proceeded on the assumption that the decisions of the participants in Libet's experiment can be modelled—as neural decision tasks typically are—in terms of an accumulator-plus-threshold mechanism: decisions are made when relevant evidence accumulated over time reaches a certain threshold. Given that in Libet's task subjects are explicitly instructed not to base their decision on any specific evidence, Schurger and colleagues proposed in this instance that the decision process amounts to simply shifting premotor activation closer to the threshold for initiation of the movement and waiting for a random threshold-crossing fluctuation in RP. Thus, Schurger and colleagues predicted the same premotor activation build-up as Libet when a movement is produced. However, whereas on Libet's post-decision interpretation of this build-up there should be no premotor activity (and hence no RPs) when no movement is produced, on Schurger and colleagues' stochastic decision model there should be continuous random fluctuations in RPs even when no movement is produced. Schurger and colleagues reasoned that it should be possible to capture these fluctuations by interrupting subjects in a Libet task with a compulsory response cue and sorting trials by their reaction times. On the assumption that the interrupted responses arise from the same decision accumulator as the self-initiated ones, and on the assumption that close-to-threshold activity reflects spontaneous fluctuations of RPs rather than mounting preparation to move building over the course of the entire trial, slow and fast reaction times should be distributed equally within trials. In their *Libetus Interruptus* task, they found, as they had predicted, that slow and fast responses to interruptions were distributed equally throughout the time span of the trial.

These results cast serious doubt on Libet's claim that the neural decision to move coincides with the onset of the RP, since spontaneous fluctuations of RPs happen all the time. There-

fore, they also cast doubt on his further claim that since RP onset precedes the urge to move by 350ms or more, conscious intentions can play no role in the initiation of the movement. If instead the neural decision to move coincides with a much later threshold-crossing event, it remains at least an open possibility that this event coincides with and constitutes the neural basis of a conscious urge to move. Schurger and colleagues take no stand on the exact relation between the conscious urge to move and their threshold-crossing event. They insist, however, that this threshold-crossing event should not be interpreted as *the* cause of the movement but rather as just one of the many factors involved in the causation of self-initiated movements. This leads me to my final point.

One conceptual problem with Libet's interpretation of his findings and also, as Dreßing points out, with most interpretations of neuroscientific experiments and a large part of the philosophical debates on mental causation and causal exclusion lies in the conception of causality that is assumed, “namely a temporal, linear, one-way causality” (Dreßing [this collection](#), p. 10). I agree with Dreßing's suggestion that a different concept of causation should be considered, one that allows for multiple causal processes to operate in parallel and to exert influence on one another. This is indeed the spirit of the dynamical model of intentions I have proposed elsewhere (Pacherie 2008). In particular, I insisted that a distal intention does not cease to exist and play a role once a corresponding proximal intention has been formed (and the same goes for proximal and motor intentions). What I suggested is that all three levels of intentions operate simultaneously, each exerting its own form of control, as well as operating together with unconscious processes. Following Dretske's lead, we can think of intentions as structuring rather than as triggering causes of action. On the dynamic hierarchical model of intentions I have proposed, we can further think of the structures set up by intentions as nested. This means that we don't need intentions to initiate actions for them to play a causal role in the production of action. This also means that the intentional online control that I argued was an

important pragmatic function of intention may be best conceived as a form of re-structuring, necessary only when the initial structuring is inadequate.

5 Conclusion

In her commentary, Dreßing suggested that the story I told about intentions should be viewed not just as a creation myth but as an attempt to give an explanation of the function of conscious intentions in the physical world. I tried to clarify exactly what I offered as merely a creation myth, namely the story given in answer to the question “*Why* do we have intentions in the first place?” and what I offered as empirical claims, namely my story as an answer to the question “*What* roles do intentions play in human agency?”

Dreßing also stresses that as an account of the roles intentions play in agency, my story has to meet both metaphysical and empirical constraints. In particular, she suggests that my claims about the role of intentions in action control makes the Causal Exclusion Problem more pressing for me than for other myth-tellers. I argued that the problem is actually equally pressing for all of us who want their views to be compatible with physicalism. I suggested that Dretske’s distinction between structuring and triggering causes and his view that mental properties should be understood as structuring causes may offer a solution to this metaphysical problem. Finally, Dreßing remarks that my claims concerning the role of conscious intentions appear to clash with certain findings from neuroscientific experiments. In response, I briefly discussed the most famous of these experiments, Libet’s RP experiments, and pointed out some of their limitations. I also questioned, together with Dreßing, the conception of causation with which these debates tend to operate, and suggested that Dretske’s distinction between structuring and triggering causes may also help to reconcile neuroscientific findings and claims about the causal roles of intentions.

References

- Anscombe, G. E. M. (1963). *Intention (second edition)*. Oxford, UK: Blackwell.
- (2008). *Faith in a hard ground: Essays on religion, philosophy and ethics (Vol. 11)*. Exeter, UK: Imprint Academic.
- Banks, B. & Pockett, S. (2007). Benjamin Libet’s work on the neuroscience of free will. In M. Velmans & S. Schneider (Eds.) *The Blackwell companion to consciousness*. London, UK: Blackwell.
- Bayne, T. & Pacherie, E. (2014). Consciousness and agency. In J. Clausen & N. Levy (Eds.) *Springer Handbook of Neuroethics* (pp. 211-230). Dordrecht, NL: Springer.
- Block, N. (1995). On a confusion about the function of consciousness. *Behavioral and Brain Sciences*, 18 (2), 227-247. [10.1017/S0140525X00038188](https://doi.org/10.1017/S0140525X00038188)
- Bratman, M. (1987). *Intention, plans, and practical reason*. Cambridge, MA: Harvard University Press.
- Chalmers, D. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2 (3), 200-219.
- Dretske, F. I. (1988). *Explaining behavior: Reasons in a world of causes*. Cambridge, MA: MIT Press.
- (2004). Psychological vs. biological explanations of behavior. *Behavior and Philosophy*, 32 (1), 167-177.
- Dreßing, A. R. (2015). Conscious intentions: Do we need a creation myth? In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences*, 8 (4), 529-566. [10.1017/S0140525X00044903](https://doi.org/10.1017/S0140525X00044903)
- Libet, B., Gleason, C. A., Wright, E. W. & Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act. *Brain*, 106 (Pt 3), 623-642.
- Mele, A. (2009). *Effective intentions: The power of conscious will*. New York, NY: Oxford University Press.
- Nahmias, E. (2002). When consciousness matters: A critical review of Daniel Wegner “the illusion of conscious will”. *Philosophical Psychology*, 15 (4), 527-541. [10.1080/0951508021000042049](https://doi.org/10.1080/0951508021000042049)
- Pacherie, E. (2008). The phenomenology of action: A conceptual framework. *Cognition*, 107 (1), 179-217. [10.1016/j.cognition.2007.09.003](https://doi.org/10.1016/j.cognition.2007.09.003)

- (2014). Can conscious agency be saved? *Topoi*, 33 (1), 33-45. [10.1007/s11245-013-9187-6](https://doi.org/10.1007/s11245-013-9187-6)
- (2015). Conscious intentions. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Robb, D. & Heil, J. (2014). Mental causation. In E. N. Zalta (Ed.) *The Stanford encyclopedia of philosophy* (spring 2014 edition).
<http://plato.stanford.edu/archives/spr2014/entries/mental-causation>
- Roskies, A. (2011). Why Libet's studies don't pose a threat to free will. In W. Sinnott-Armstrong & L. Nadel (Eds.) *Conscious Will and Responsibility* (pp. 11-22). New York, NY: Oxford University Press.
- Schurger, A., Sitt, J. D. & Dehaene, S. (2012). An accumulator model for spontaneous neural activity prior to self-initiated movement. *Proceedings of the National Academy of Sciences*, 109 (42), E2904-E2913. [10.1073/pnas.1210467109](https://doi.org/10.1073/pnas.1210467109)
- Slors, M. (2015). Conscious intending as self-programming. *Philosophical Psychology*, 28 (1), 94-113. [10.1080/09515089.2013.803922](https://doi.org/10.1080/09515089.2013.803922)
- Velleman, D. (2007). What good is a will? In A. Leist & H. Baumann (Eds.) *Action in Context* (pp. 193-215). Berlin, GER: de Gruyter.
- Wu, W. (2011). Confronting many-many problems: Attention and agentic control. *Noûs*, 45 (1), 50-76. [10.1111/j.1468-0068.2010.00804.x](https://doi.org/10.1111/j.1468-0068.2010.00804.x)

Naturalizing Metaethics

Jesse Prinz

Decades ago, it was suggested that epistemology could be naturalized, meaning, roughly, that it could be treated as an empirically-informed psychological inquiry. In more recent years, there has been a concerted effort to naturalize ethics, with a focus on questions in moral psychology, and occasional normative ethics. Less effort has been put into the naturalization of metaethics: the study of what, if anything, makes moral judgments true. The discussion presents a systematic overview of core questions in metaethics, and argues that each of these can be illuminated by psychological research. These include questions about realism, expressivism, error theory, and relativism. Metaethics is beholden to moral psychology, and moral psychology can be studied empirically. The primary goal is to establish empirical tractability, but, in so doing, the paper also takes a provisional stance on core questions, defending a view that is relativist, subjective, and emotionally grounded.

Keywords

Error theory | Expressivism | Metaethics | Moral realism | Naturalism | Relativism | Sentimentalism

Author

Jesse Prinz

jesse@subcortex.com

City University of New York
New York, NY, U.S.A.

Commentator

Yann Wilhelm

ywilhelm@students.uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

Moral philosophy has taken an empirical turn, with experimental results being brought to bear on core questions in moral psychology (e.g., is altruism motivated by empathy?) and normative ethics (e.g., how plausible are the presuppositions of virtue theory?). Some of the recent empirical work also bears on core questions in metaethics. Metaethical questions are varied, but they broadly concern the foundations of moral judgments. What is the basis of such judgments? What, if anything, could render them true? Here I will argue that these questions can be empirically addressed, and longstanding debates between leading metaethical theories may ultimately be settled experimentally. I will describe empirical results that bear on core metaethical questions. I

will not present these results in detail here. My goal is programmatic: I seek to establish the empirical tractability of metaethics. Some of the experiments I describe are exploratory pilot studies, presented in an effort to motivate more research. Even with such preliminary results, we will see that some metaethical theories already enjoy greater empirical support than others. I will argue that the best-supported theory at this stage of inquiry is a form of relativist sentimentalism. Defending this position is subsidiary to my primary goal of advertising the value of empirical methods in metaethical theorizing. There has already been an empirical turn in ethics, but metaethics has been less explicitly targeted by these new approaches.

Talking about “an empirical turn” clearly alludes to another turn in the recent history of

philosophy: the linguistic turn. When philosophers turned their attention to language, there was an effort to recast philosophical problems as linguistic in nature. A new set of technical tools was brought into the field: formal semantics. Logic has been part of philosophy historically, but after the linguistic turn it was perceived to be an essential component of philosophical training. Just as formal semantics increased philosophical precision with the linguistic turn, empirical methods have dramatically augmented our tool chest, and stubborn debates may begin to give way. The empirical turn is as momentous as the linguistic turn, and perhaps even more so. Formal semantics allowed us to articulate differences between theories, and empirical methods provide new opportunities for theory confirmation. Neither turn rendered traditional approaches to philosophy idle, but rather supplemented them. Within metaethics, this supplementation may offer the best hope of settling which competing theories are true.

In calling for a naturalist metaethics, it is important to avoid confusion with two other views. “Naturalism” is sometimes construed as a metaphysical thesis, and also sometimes as a semantic thesis. Metaphysically, “naturalism” refers to the view that everything that exists belongs to the natural world, as opposed to the non-natural, or supernatural world. This is sometimes presented as a synonym for physicalism, which can be defined as the view that the world described by the physical sciences is complete, in that any physical duplicate of this world would be a duplicate simpliciter. The causal closure of the physical world and the success of physical science are taken as evidence for this metaphysical view. Semantic naturalism attempts to reductively analyze concepts from one domain in terms of another, which is considered more likely to be natural in a metaphysical sense. In philosophy of mind, this might involve defining psychological concepts in neural or causal terms, while in ethics it might involve defining moral properties in terms of psychological, logical, or social terms (such as hedonic states, principles of reason, or social contracts). Here I will be concerned with methodological naturalism, which has recent roots in the work of

W.V.O. Quine, who grew skeptical about philosophizing through linguistic analysis, and emphasized the empirical revisability of philosophical claims (1969). Quine drew on the methods of John Dewey, and insisted that “knowledge, mind, and meaning [...] are to be studied in the same empirical spirit that animates natural science” (1969, p. 26). More succinctly, methodological naturalism can be defined as follows:

Methodological naturalism =_{Df} the view that we should study a domain using empirical methods.

This is the kind of naturalism that has long been advocated, but too rarely followed, in the domain of epistemology (Kornblith 1985). Neither metaphysical nor semantic naturalism are equivalent to methodological naturalism. Metaphysical naturalism is a view about what exists, not about how to study it. Indeed, some non-naturalists in this metaphysical sense believe that empirical methods can be used to study non-physical or supernatural entities. Semantic naturalism is a view about how to state theories (viz., in reductionist terms), but practitioners have rarely used empirical science in defense of such theories (consider so-called naturalistic semantics). Methodological naturalism has been deployed in discussions of both first-order ethics (e.g., Brandt 1959; Flanagan 1991; Doris 1998; Greene 2007) and in metaethics (e.g., Railton 1993; Prinz 2007b). As Railton points out, a naturalist methodology could result in a reductionist theory of morality, but it need not (see also Boyd 1988). In principle, science could support traditional intuitionism, which is not naturalistic in either of these other senses.

1.1 Methodological preamble

Philosophy has always been methodologically pluralistic. Some use intuitions to arrive at necessary and sufficient conditions for the application of concepts (e.g., Plato’s early dialogues). Some try to systematize and revise a large set of beliefs using reflective equilibrium (e.g., Rawls on justice). Some use transcendental ar-

guments to figure out preconditions for thought and action (e.g., Kant). Some use aphorisms or stories to reveal facts about ourselves or to envision possible alternatives (e.g., Nietzsche and the existentialist tradition). Some propose historical analyses of prevailing institutions and values (e.g., Hobbes, Rousseau, and Foucault). Some disclose hidden social forces that buffer prevailing categories (e.g., Marilyn Frye on gender). Some analyze case studies (e.g., Kuhn), probe the structure of experience (e.g., Husserl), or propose formalizations (e.g., Frege). These and other methods suggest that philosophy is a many-splendored thing, and among its many forms one can also find the deployment of empirical results. Examples include Descartes and James on the emotions, Merleau-Ponty on embodiment, and Wittgenstein on aspect perception. Empirical observations have often guided philosophical inquiry. Locke was inspired by corpuscular physics, Marx took solace in Darwinian biology, and Carnap incorporated ideas from behaviorism.

The term “empirical” is used in different ways. In its broadest application, it refers to observational methods. Observation can include an examination of the world, both inner and outer, with and without special instruments. Even introspection can be regarded as a form of observation, as the etymology of the term suggests, and in this sense, the introspection of intuitions is an observational method. Philosophers who use intuitions in theory-construction can be characterised as doing something empirical in this broad sense. Linguistics has used such intuitions to construct syntactic theories, and few would deny that syntax is an empirical field. But the term “empirical” is also used more narrowly to refer to the use of scientific methods, which involve the design of repeatable observation procedures, and the quantification and mathematical analysis of data. The empirical turn in philosophy has been marked by a dramatic increase in the use of scientific results.

Many philosophers have long held a positive attitude toward science, but the frequent use of scientific results (outside of the philosophies of science) is a recent phenomenon. It became popular in naturalized epistemology, which

draws on the psychology of decision-making, and philosophy of mind, which has drawn on psychology, computer science, and artificial intelligence. Over the last decade, empirical methods have also become widely used, and widely contested, in ethics.

The resistance to empirical methods in ethics is often chalked up to the fact that ethics is a normative domain, and empirical methods provide descriptive results. This can only be part of the story, however, as there has been little uptake of empirical methods in metaethics. Metaethics is a descriptive domain; it does not tell us how to act morally, but rather explores the semantic commitments and metaphysical foundations of such claims. I suspect the reason for resistance is less interesting and more sociological. Psychology is a young profession, which grew out of philosophy and physiology but then acquired its own institutional standing in the academy, and it has had to fiercely guard its status as a science by distancing itself from the humanities. Meanwhile, philosophy underwent an analytic turn, which led to a preoccupation with conceptual analysis, and an anxiety about psychologism. On this vision, the field began to model itself on logic or mathematics, which were, in turn, taken to be *a priori* domains. I think this is a fundamental mistake. In many domains, the concepts that matter most are grounded in human usage, not in a transcendental realm like (allegedly) mathematics. The arbiters of conceptual truth include both the inferences we are inclined to draw and our linguistic behavior, both of which can be studied empirically. I will not argue directly against *a priori* approaches, but will instead make an empirical case, or better yet an invitation, by attempting to illustrate how empirical findings make contact with traditional philosophical questions in metaethics.

One manifestation of the empirical turn has been the rise of experimental philosophy. This term most often refers to the work of philosophers who conduct studies that probe people’s intuitions about philosophical thought experiments. Strictly speaking, much of this work is not experimental, since the term “experiment” is often reserved in psychological re-

search for studies in which researchers attempt to manipulate the mental states of their participants—experimental conditions are compared against control conditions. Experimental philosophy often explores standing intuitions, rather than the factors that influence those intuitions (e.g., [Mikhail 2002](#)). For example, some trolley studies simply poll opinions about the permissibility of certain actions. That is a survey rather than an experiment. One can use survey methods to conduct experiments, however. For example, one could conduct a trolley study in which some vignettes use evocative language in an effort to manipulate participants' emotions. Few experimental philosophy studies do anything like this. Most ask for opinions without manipulating psychological states. Thus, experimental philosophy characteristically examines the *content* of people's concepts and beliefs, rather than the underlying psychological processes. In this sense, experimental philosophy is an extension of conceptual analysis. For those interested in underlying processes, it can, to this extent, be of limited interest. Some experimental philosophy has also been criticized for failing to meet standards of reliable behavioral research ([Woolfolk 2013](#)). That said, conceptual questions are often important for philosophical theorizing, and methodological problems with experimental philosophy can be addressed by conducting better and better experiments. Often the first efforts (including much of the work I will describe below) are best regarded as analagous to pilot studies, in need of refinement but successful enough to warrant more careful investigation. In addition, many philosophers draw on (and increasingly conduct) studies that qualify as genuine experiments and meet the standards of good social science. There is a long tradition of philosophers using research published in social science journals to defend philosophical positions. For those who find paradigm cases of experimental philosophy too limiting (because they are based on conceptual intuitions or fail to meet certain standards), there are many other empirical results that can provide illumination. The term "empirical philosophy" can be used as a broader label to cover both opinion polls and experimental manipula-

tions. As I use the term, it refers to the use of scientific results, whether obtained by a philosopher or not, to address philosophical questions. The empirical turn should not be dismissed as philosophy-through-opinion-polls; it is a multi-pronged effort advance philosophical debates by drawing upon observational methods of any kind.

The motivations for the empirical turn are varied, but the most general impetus is the belief that some questions cannot be resolved by more traditional philosophical methods. For example, philosophers interested in the physical basis of consciousness cannot rely on introspection or on an analysis of the concept "conscious." And even those interested in analysis of concepts have worried about the limits of introspection. There are basically three different theories of what concepts are: Platonic entities, emergent features of linguistic practice, or psychological states. None of these can be completely investigated by introspection. Even psychological states can be difficult to introspect, because much mental activity is unconscious, and because introspection may be prone to error and bias. Moreover, even if a philosopher could perfectly introspect her own concepts, she would not know thereby that others shared the same concepts, and this would greatly limit the scope of her theories. Some experimental philosophers have argued that philosophers' intuitions are not shared by laypeople. When philosophers and laypeople do agree on intuitions, there is still no guarantee that these accurately reflect reality. For example, most people (at least in the West) find it intuitively plausible that human action derives from character traits, but some empirical philosophers (most notably Owen Flanagan, John Doris, and Gilbert Harman) have drawn attention to psychological research that challenges this assumption.

Traditional and empirical approaches to philosophy are sometimes placed in opposition, but they can also be regarded as interdependent. On one division of labor, traditional methods are used to pose questions and to devise theories that might answer those questions. Empirical methods can then be used to test these theories. This is an over-simplification, of course, because observa-

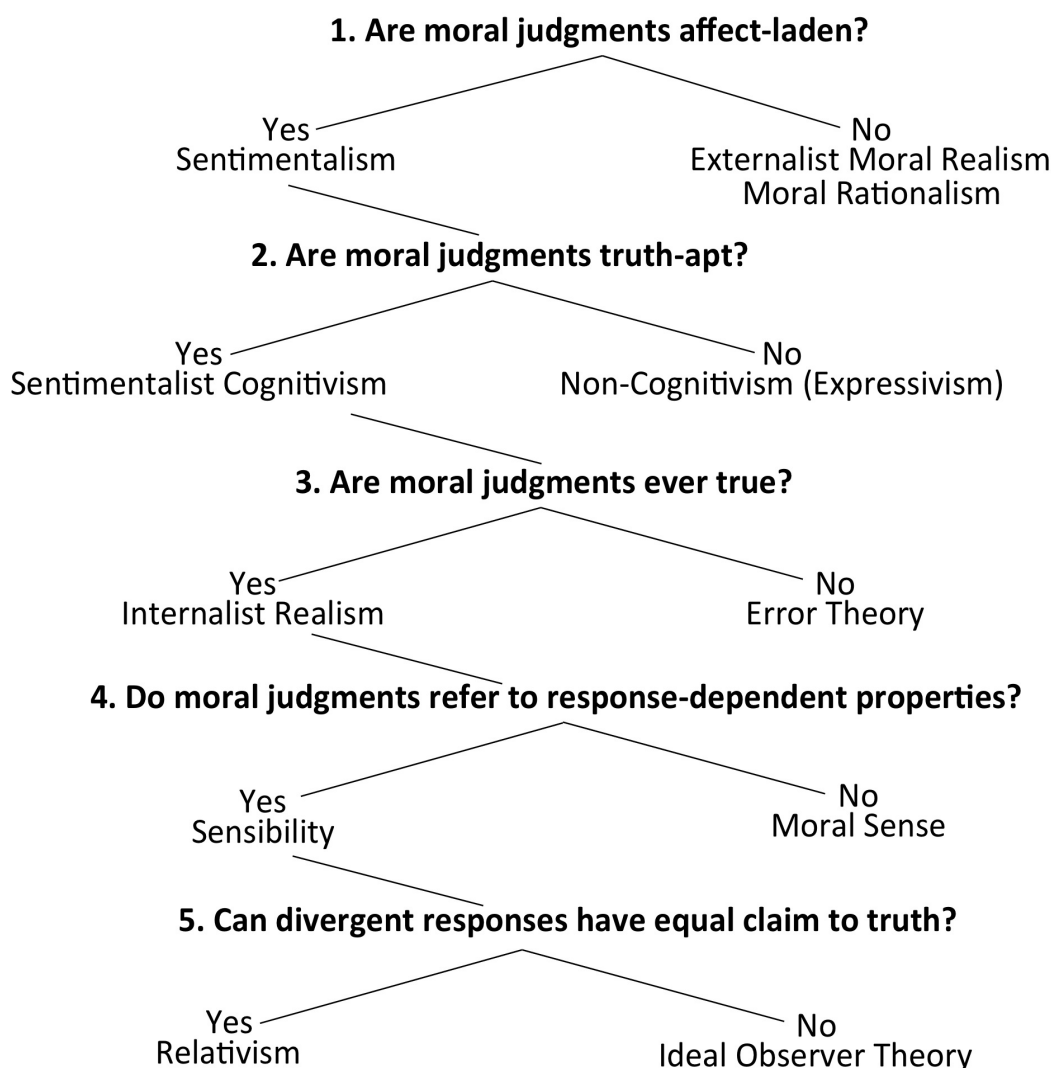


Figure 1: A Metaethics decision-tree

tions can inspire theories, and traditional methods can sometimes refute theories (Gettier cases are a parade example in epistemology), but the proposed division of labor is a decent approximation. Traditional methods have limited testing power because theoretical posits are often difficult to directly observe, and empirical methods have limited power in constructing theories, because theories outstrip evidence. In what follows I will test theories derived from philosophical reflection against the tribunal of empirical evidence.

1.2 A roadmap to metaethics

Let us turn now to the focus of discussion: metaethics. Metaethical questions concern the nature of the moral domain. Metaethicists ask: what kinds of things are we talking about when

we make moral judgements? Put differently, metaethics concerns the truthmakers of moral judgements: what kinds of facts, if any, make moral judgements true? That is a metaphysical question but it is normally approached semantically by exploring what we are semantically committed to when we make moral judgements. Metaethics differs from first-order ethics, which concerns the content, derivation, and application of such judgements.

There are many different metaethical theories, and a complete survey here would be impossible. I will focus on major theories that have emerged over the last two hundred and fifty years, with emphasis on proposals that dominated discussion in the twentieth century. To be clear from the outset, my goal is not to consider specific theories that have been ad-

vanced by currently active authors in metaethics. Rather, I will survey broader classes of theories that have been around for some time (decades or centuries) in an effort to establish the relevance of empirical work. An adequate examination of any recent theory would require a narrower focus than I am after here, since each theory makes empirical commitments, if at all, in different places.

To facilitate discussion, I will map out the theories of interest using a decision tree (Figure 1). The tree could easily be arranged differently. Almost any branch could be the starting place, with other nodes occurring higher or lower than they are in this rendition. As we will see in a moment, I begin with a question about “affect” or emotions. This may seem odd to some contemporary metaethicists. Some contemporary metaethicists discuss emotions (such as [Alan Gibbard 1990](#), and [Simon Blackburn 1998](#)), but others do not (for example, emotions are discussed less among moral realists). Historically, however, emotions have been a central focus in metaethics. British moralists, who advanced many of the questions that continue to drive the subfield, often begin their analyses with a discussion of moral sentiments. Indeed, the most famous controversy in metaethics before the twentieth century is probably the dispute between British sentimentalism and Kantian rationalism. Even in the twentieth century, some of the most discussed debates concern emotions, such as the debate between emotivists and their opponents. Moreover, the recent empirical turn was triggered, in part, by discoveries linking emotions to moral judgement. So this starting point has considerable historical depth and great relevance to the methodological sea-change that I am interested in here. That said, I don’t intend this tree to be anything like a complete map of meta-ethics. One could begin elsewhere and branch out in further directions (I expand the tree leftward, but interesting questions also come up on the right). Though incomplete, the nodes of this tree encompass much of what one might cover in an introduction to metaethics in the Anglophone analytic tradition.

The first question in the metaethics decision tree is, I note, a question about emotions. More precisely, we can ask: are moral judgments affect-laden? The term “affect” is used instead of “emotion” here, because it is broader. I intend the term to cover any conative state, such as a preference, desire, or pro-attitude. For most of this discussion, I will focus on emotions rather than these other affective constructs, because emotions are implicated in the empirical work I will be considering.

The other key term in question 1 is “moral judgments.” By “moral judgments” I mean atomic judgments using thin moral concepts, such as *wrong*, *bad*, or *immoral*. The judgment expressed by “Shoplifting is wrong” would be an example. There are many other judgments that arise in moral contexts, including judgments containing thick concepts, such as *cruel* or *unjust*. One can also ask whether these are affect-laden. On one analysis, thick concepts are hybrids that have both a descriptive and an evaluative component, the latter of which may implicate the emotions. For the sake of simplicity I will ignore that debate here.

Notice that judgements are not sentences but rather the thoughts that sentences express. To propose that such thoughts are affect-laden is to say that each token instance involves an emotion or other conative state. There are different forms of “involvement” that have been discussed in metaethics. On some theories, moral judgments contain conative states as constituent parts. This was the view of Francis Hutcheson, David Hume, and some other British moralists. One might weaken this by saying that moral judgments do not contain emotions, but refer to them. In this vein, John McDowell, David Wiggins, and Alan Gibbard suggest that moral judgments reflect the conviction or norm that it would be warranted to feel certain emotions. Both of these approaches have gone under the heading “sentimentalism,” with the prefix “neo-” for the views that say the link between moral judgments and emotions is second-order. Here is a definition:

Sentimentalism =_{Df} Moral judgments essentially involve affective states, such as emotions, in one of two ways: such states are constituent parts of moral judgments (traditional sentimentalism); or moral judgments are judgments about the appropriateness of such states (neo-sentimentalism).

Those who deny that moral judgments are affect-laden fall into different categories, but two of the most important metaethical theories of this kind are externalist moral realism and (some forms of) rationalism. Moral realists say that there are moral facts, which is to say that some states of affairs are truly right or wrong (cf. [Sayre-McCord 1988](#)). Externalist moral realists add a further requirement, namely mind-independence:

Externalism moral realism =_{Df} There are moral facts and these obtain independently of our recognition of them.

If moral facts are mind-independent, it also follows that we can come to know them without being moved by them. Like scientific facts, we can know that they obtain without feeling any way towards them. Cornell realists, some intuitionists, and many divine command theorists fall into this category. Moral rationalism is a view about how the epistemology or normative status of moral truths:

Moral rationalism =_{Df} Moral truths can be discovered and justified through a purely rational decision procedure.

[Kant \(1797\)](#) is traditionally read this way, though he also claimed that moral judgments involve moral feelings.

The remainder of my metaethics decision tree concerns those who think that moral judgments are affect-laden. Among those who say that moral judgments essentially involve conative states, there is a divide between those who think that moral judgments are nevertheless truth-apt and those who deny this. This is the second division of the tree. A judgment is truth-

apt if it is the kind of thing that can be evaluated as true or false. Some affect-laden judgments may turn out to have a merely expressive function. If I say, “Disco sucks!” I may not be attempting to represent a fact, but merely expressing how I feel. Expressivists follow this analogy:

Expressivism =_{Df} Moral assertions express mere feelings or non-assertoric attitudes, and do not purport to convey facts.

Charles Stevenson and A. J. Ayer are credited with devising the emotivist theory of morality, which is the simplest theory of this kind. A more sophisticated variant has been developed by Simon Blackburn, who proposes that moral judgments aspire for quasi-truth, but not truth, and thus an ontologically neutral stand-in—which can explain why moral judgments have an assertoric form. Alan Gibbard says that moral judgments do not directly express feelings, as emotivists claim, but rather express the acceptance of norms according to which feelings such as anger and guilt would be appropriate. All these theories have been broadly classified as expressivist.

Those who say that moral judgments are affect-laden and truth-apt need not deny that moral judgments are expressive, but they insist that they more than express feelings; they assert facts. If so, moral judgments can be true or false. Subjectivism falls into this camp:

Subjectivism =_{Df} the truth of the judgment that something is morally good or bad depends on the feelings or other subjective states of someone who makes that judgment.

For instance, one might propose that “killing is wrong” means “I disapprove of killing.” That judgment is true, if the speaker disapproves of killing, and false otherwise. As we will see, there are also more sophisticated forms of subjectivism. Subjectivists are internalist moral realists: they believe in moral facts, but they deny that those facts obtain independently of our attitudes.

The term “cognitivism” has been used to refer to any view on which moral judgments are truth-apt, which is to say they can be assessed for truth. Expressivists are non-cognitivists, and both subjectivists and external moral realists are cognitivists. One could also have a cognitivist theory and nevertheless insist that all moral judgments are false. This would be an error theory.

Error theory =_{Def} Moral judgments are truth-apt, but they are never true.

The most famous error theory comes from J. L. Mackie. Mackie argues that moral judgments are incoherent. On the one hand, they presuppose that moral facts are objective, which is to say mind-independent. On the other hand, moral judgments presuppose that moral facts are action guiding, and that suggests that they directly motivate us. This suggests that moral judgments must be affect-laden, or otherwise dependent on our subjective states. Since nothing can be both objective and subjective, moral judgments can never be true. Opponents of the error theory deny this and insist that some moral judgments are true. They are, in this sense, moral realists. Moral realists who also claim that moral judgments are affect-laden must take Mackie’s challenge head on, showing that truth is compatible with being action-guiding.

Such sentimentalist realists face an immediate question. They can accept Mackie’s claim that moral judgments represent objective properties, and find some way to circumvent the incoherence, or they could say that moral judgments refer to properties that are subjective, or response-dependent. The first option might seem untenable, since it accepts that moral judgments are both objective and subjective, an apparent contradiction. But the contradiction can be mitigated by distinguishing between sense and reference. One might say that moral concepts have affect-laden senses—that is, we grasp them by means of feeling—and objective referents. Consider, for example, Kant’s aesthetics, according to which beauty consists in a kind of purposeful purposelessness that causes a free-

play of the understanding, which results in aesthetic pleasure. A work may have purposeful purposelessness without our recognizing that this is so, but when we recognize it, we feel a certain way. Within ethics, Francis Hutcheson may have held a view that was objectivist and subjectivist in just this way. He suggests that moral facts are established by divine command, but God has furnished us with a moral sense, and that sense works by means of the emotions; when we see objectively bad actions, we feel disapprobation. This has been called a moral sense theory, because it treats our moral passions as a kind of sensory capacity that picks up on real moral facts.

In contrast to this view, one might argue that moral facts are not objective, as Mackie has maintained, but rather are dependent on our responses. This need not imply that moral judgments are mere expressions of feeling; one might say instead that moral judgments refer to response-dependent properties. The idea of response-dependent properties derives from John Locke’s notion of secondary qualities. Primary qualities, such as shape, for Locke, exist independently of being perceived, whereas secondary qualities consist in the power that certain things have to cause responses in us. Colors, for Locke, are not out there in the world, but consist in the fact that objects cause certain sensations in us. The moral analogue of this view has been called the sensibility theory, and its adherents include John McDowell, David Wiggins, and David McNaughton. They resist the causal language found in Locke’s theory of colors, but say something close:

Sensibility theory =_{Def} moral properties are those that warrant moral emotions.

Strictly speaking, the sensibility theory is a form of subjectivism, since it says that moral judgments refer to subjective properties (the property of warranting moral emotions), but the notion of warrant allows these theorists to avoid a pitfall or simple subjectivism. For a simple subjectivist it makes no sense to wonder whether something that I disapprove of is really wrong, but for the sensibility theories I can en-

certain such doubt because I can wonder whether an event really warrants what I happen to feel. The notion of warrant here is not unproblematic, and it often goes unanalyzed. There is one notable exception, however, and that is the ideal observer theory (Firth 1952; Brandt 1959):

Ideal observer theory =_{Def} The morally good or bad is that which an observer would regard as good or bad under ideal circumstances.

Such circumstances might involve acquiring the status of a moral sage (or consulting a moral sage), as on some virtue theoretic theories, or an ideal version of myself (Smith 1994). Ideal observer theorists are committed to response-dependence; they say that responses determine moral truth, and they further require that those responses come from certain kinds of epistemic agents.

Ideal observer theories offer a negative answer to the final question in the metaethics decision tree. They specify conditions of ideal observation in order to find an authoritative set of responses among a diversity of moral opinions. The hope is that one set of judgments will emerge as epistemically superior to all others; on this view, all moral judges converge under ideal conditions. Here, moral truths work out to be universal. This, of course, is a controversial claim. Suppose we define ideal observers as those who are free from bias, aware of pertinent non-moral facts, and reasoning carefully. It could turn out that, two such observers could still disagree on moral matters. This prognosis leads toward the view that there is no way to arrive at moral consensus. Those who think that moral judgments are rendered true by a judge's response but deny consensus under optimal epistemic conditions end up saying that moral judgments are relative. This view can be stated as follows:

Metaethical relativism =_{Def} Two judgments expressed using tokens of the same word types, and grasped by tokens of the same mental attitude types can have different

truth-values if they are made by different observers.

I will now try to show that each question on the decision tree can be empirically illuminated. Some of the empirical results that I will present come from unpublished, exploratory studies. My goal here is not a detailed documentation of scientific findings, but rather to establish, by means of example, ways in which empirical methods might be brought to bear on the foregoing questions. The hope is that the studies described here might be taken up by others and improved upon.

2 Empirical resolutions to metaethical debates

2.1 Sentimentalism vs. rationalism and externalism

Let's begin with the first question on the metaethics decision tree: are moral judgments affect-laden? No question in ethics has received more empirical attention than this. Dozens of studies have attempted to determine whether emotions play a central role in morality, and the evidence has consistently shown that they do. Let me begin with an unpublished study of my own and then offer a brief review of the empirical literature.

To begin with, let's consider folk intuitions. Do ordinary people use emotions as evidence when attributing moral judgments? To test this, I conducted a simple vignette study, which pitted emotions against verbal testimony. A group of college undergraduates taking an introductory-level philosophy class responded to the following probe:

Fred belongs to a fraternity and his brothers in the fraternity sometimes smoke marijuana. Fred insists that he thinks it's morally acceptable to smoke marijuana. He says, "You guys are not doing anything wrong when you smoke." But Fred also feels disgusted with his frat brothers when he sees them smoking. One day, to prove that he thinks smoking is okay, he smokes marijuana himself. Afterwards, he feels incredibly ashamed about smoking the drug.

Which of the following seems more likely:

1. Fred says he morally approves of marijuana smoking, but in reality he thinks it is morally wrong.
2. Fred feels badly about smoking marijuana, but in reality he thinks it is morally acceptable.

In my small sample, 68.4% chose answer 1, suggesting that the majority of them take emotions as evidence for moral values, even when that directly contradicts self-report. This suggests that many people take emotions to be a sufficient evidence for attribution moral attitudes. An even more dramatic result was obtained when another twenty participants assessed this scenario:

Frank belongs to a fraternity and his brothers in the fraternity sometimes smoke marijuana. Frank insists that their actions are morally unacceptable. He says, “You guys are doing something wrong when you smoke.” But Frank does not feel any anger or disgust when he sees his frat brothers smoking. One day, when they are not around, he smokes marijuana himself. Afterwards, he doesn’t feel any shame about smoking the drug.

Which of the following seems more likely:

1. Frank says he morally opposes marijuana smoking, but in reality he thinks it is morally acceptable.
2. Frank doesn’t feel badly about smoking marijuana, but in reality he thinks it is morally wrong.

Here, 89.5% of participants chose response 1, indicating that they take emotions to be necessary for the attribution of moral attitudes. Absent the right feelings, verbal testimony is treated as an unreliable indicator of a person’s values.

This study has at least four serious limitations: people may not trust self-reports; the results were far from unanimous; it fails to distinguish evidence for moral attitudes and essence of moral attitudes; and folk beliefs about moral judgments may be wrong. To get around these

limitations we must move beyond experimental philosophy, and look for more direct evidence that emotions actually are sufficient and necessary for moral judgments. But the study is still revealing, because it shows that emotions are used as evidence in moral attribution. Most participants make attributions that fall in line with sentimentalism.

To show that emotions actually do contribute to moral cognition, we can look at three kinds of evidence: cognitive neuroscience, behavioral psychology, and pathology. In each domain, sentimentalism finds support. There have now been dozens of neuroimaging studies on moral judgment tasks, and every one of them, to my knowledge, shows an increase in activation in brain structures associated with emotion, when moral decisions are compared to non-moral decisions. Key structures include the posterior cingulate, temporal pole, orbitofrontal cortex, and ventromedial prefrontal cortex. There are only two groups of studies that even appear to depart from this pattern. [Joshua Greene et al. \(2001\)](#) report that emotions play more of a role in deontological judgments than in consequentialist judgments, but their data show that, as compared to non-moral judgments, emotions are involved in both (see their Figure 1). Moreover, Greene et al. use moral dilemmas in which the common denominator is saving lives—they manipulate the nature of the harm necessary in order to save five people in danger. Thus, each moral judgment condition presumably elicits the judgment that it would be good to help people in need. This positive moral judgment may be emotionally grounded, but the neuroimaging method subtracts away this emotional information, because it is present in each scenario, and imaging results of this kind report only contrasts between different conditions. Thus, a major dimension of moral emotions may be systematically concealed by the method. The other study that fails to show an increase in emotional responses during moral judgment is one condition in a series of imagining experiments performed by [Jana Borg et al. \(2006\)](#). But, in that condition, a moral scenario is compared to a scenario about an encroaching fire that threatens one’s property, and it is un-

surprising that moral judgments produce less of an emotional response than a case of personal loss.

Brain science resoundingly links moral judgment to emotion, but the method is correlational. Moral rationalists and externalists could concede that moral judgments excite emotional responses, while denying that these are the basis of moral judgment. Imagine the following view: we use reason to arrive at moral judgments, but morality matters to us, so when we arrive at those judgments emotions normally kick in. By analogy, reason might be used to determine that certain life activities (smoking, high fat diets, sleep deprivation) are harmful, and, upon drawing that reason-based conclusion, we tend to experience corresponding emotions, such as anxiety when contemplating lighting a cigarette. Neuroimaging results showing responses to cigarettes might confirm this, showing emotion areas active when cigarettes are seen, but that wouldn't refute a rationalist theory of how we arrive at the judgment that cigarettes are dangerous.

To adjudicate between the thesis that emotions are constitutive of moral judgments, as opposed to mere consequences, we need behavioral evidence. Numerous studies now establish a causal link between emotion and moral judgment. When emotions are induced, they influence how good or bad things seem. Induction methods have been widely varied: hypnosis, dirt, film clips, autobiographical recall, and smells. In one recent study, Kendal Eskine, Natalie Kacinik, and I induced bitterness by giving people a bitter beverage and found that moral judgments became more severe ([Eskine et al. 2011](#)). In other recent studies Angelika Seidel and I use sound clips to induce specific emotions, and we have shown that different emotions have different moral effects: anger induces more stringent wrongness judgments about crimes against persons; disgust induces greater stringency on crimes against nature (such as cannibalism); and happiness induces stronger judgments that it is both good and compulsory to help the needy ([Seidel & Prinz 2013a, 2013b](#)). There is also evidence that we feel different emotions when judging our own actions

than when judging others. When another person commits a crime against nature, we tend to feel disgust, but when we perform an act deemed by others to be unnatural, the most common response seems to be shame. Conversely, when others commit crimes against persons, we feel angry, but guilt is the natural response when we perform such acts ourselves. To test this hypothesis, I conducted a forced-choice study in which a group of college undergraduates had to pick guilt or shame in response to mildly "unnatural" acts ("Suppose your roommate catches you masturbating"), and mildly harmful acts ("Suppose you take something from someone and never return it"). 80% chose shame for the first case, and over 90% picked guilt for the second.

Such findings demonstrate that different emotions play different roles. I mentioned three distinctions that are currently receiving empirical attention: the split between positive and negative emotions (praise and blame), between two kinds of blame (crimes against nature and crimes against persons), and between self- and other-directed blame. The self/other distinction may be particularly important because it helps us see how moral emotions differ from their non-moral analogues. Anger (or at least irritation) and disgust can both occur in non-moral contexts, but they take on a moral cast, I submit, when paired with dispositions to feel guilt and shame, respectively. If I find eating insects physically revolting, I will experience disgust when I see others eat insects, and disgust when I inadvertently eat them myself. But if I found insect eating immoral, it would not be disgust that I experience in the first-person case, but shame. This feeling of shame would motivate me to make amends for my actions, or to conceal my wrongdoing from others, not simply to repel the unwanted food from my body. The self-directed emotions round out the punitive cast of our moral attitudes. We see morally bad acts as not just worth aggressing against, but as worthy of apology. This need not be a second-order belief. Rather it is implicit in the fact that moralized behaviors carry emotional dispositions toward self

and other that together promote a punitive attitude: a disposition to issue and submit to punishment.

Putting this together, I propose that standing moral values (the values that a given individual has for an extended period of time) consist in dispositions to feel the self- and other-directed emotions that I have been discussing. Such an emotional disposition can be called a sentiment. On any given occasion when a standing value becomes active in thought—i.e., when a moral judgment is made—these dispositions result, all else being equal, in an emotional state. The emotion that is felt depends on who is doing what to whom. For example, if I recall a situation in which I hurt someone's feelings, I will have a feeling of guilt regarding that event, because a person was harmed and I was the culprit. This feeling of guilt toward an event constitutes my judgment that the action was wrong, and I gain introspective access to this judgment by feeling guilt well up inside me. If this is right, then emotions are not merely effects of moral judgments, but essential components of them.

Against this picture, one might object that emotions are merely a heuristic that can be used in certain circumstances, but not strictly necessary for making moral judgment. Following the analogy mentioned before, anxiety might be used as a heuristic when deciding whether to smoke, but the judgment that smoking is dangerous does not depend on fear, and was initially arrived at by the light of reason.

To establish that emotions are not merely helpful heuristics, one must see what happens when emotions are reduced or eliminated. To look into this, [Eskine \(2011\)](#) gave people the bitter taste manipulation and then warned them not to let the feelings caused by that beverage interfere with the moral judgments. In this condition, he found that moral judgments were considerably less severe than a control condition, suggesting that, when we ignore emotions, it is harder to see things as wrong. The finding indicates, in other words, that moral judgments subside when emotions are absent. The study cannot confirm this strong claim, however, because people cannot suppress emo-

tions completely. More powerful evidence comes from the clinical populations who suffer from emotional deficits. For example, psychopaths, who suffer from deficit in guilt and other negative emotions, notoriously fail to appreciate what is wrong with their actions ([Hare 1993](#)). Similarly, people with Huntington's disease, which impairs disgust, show high incidence of paraphelias, suggesting that they cease to see deviant sexual behavior as wrong ([Schmidt & Bonelli 2008](#)). [Kramer \(1993, p. 278\)](#) argues that anti-depressants can flatten affect in a way that results in a "loss of moral sensibility." There is also a positive relationship between alexithymia and Machiavellianism, suggesting that a reduction in emotional competence may act in ways that are more instrumental than moral ([Wastell & Booth 2003](#)). For better or worse, there is no clinical condition in which all emotions are absent and behavioral function remains, but these findings suggest that selective or global dampening of the emotions leads to corresponding deficits in moral judgment. That is, people with diminished emotions seem to be insensitive to corresponding parts of the moral domain, suggesting that they may not be forming moral judgments.

The evidence summarized here suggests that emotions arise when we make moral judgments, that emotions are consulted when reporting such judgments, and that moral judgments are impaired when emotions are unavailable. Some of this evidence is preliminary, but, for present purposes, let's assume that the findings hold up to further and more stringent testing. By inference to the best explanation, such findings suggest that emotions are components of moral judgments. The idea is that, when people say something is morally bad, the thought they are expressing on that occasion consists of a negative emotion directed towards the thing judged bad. Emotions, on this view, function like predicates in thought. That is what traditionally sentimentalists, such as Hume, seem to have maintained. Hume thought ideas—the components of thoughts—were stored copies of impressions, and the idea of moral badness consisted in a stored copy of the impression of disapprobation.

Traditional sentimentalism, which says that emotions (or sentiments) are actually components of moral judgments, differs conspicuously from neo-sentimentalism. Neo-sentimentalists theories say that moral judgments are judgments about the appropriateness of emotions. These theories do not straightforwardly predict that emotions come on line when we make moral judgments, nor that a reduction in emotions should interfere with our ability to moralize. Instead, they predict that people will think about emotions when they make moral judgments. Correlatively, they also predict that people with limited metacognitive abilities will lose their ability to make moral judgments; this is not the case (Nichols 2008). Thus, given the current state of evidence, traditional sentimentalism outperforms neo-sentimentalism empirically. Traditional sentimentalism predicts a robust pattern of empirical findings.

Rationalists and externalist moral realists might balk at this point and say that the empirical evidence lacks the adequate modal strength to support sentimentalism. The evidence shows that emotions are often consulted when making moral judgments, but this leaves open the possibility that we might also make moral judgments dispassionately under circumstances that have not yet been empirically explored. So stated, this objection is just an expression of faith. It suffers from both conceptual and empirical weaknesses. Conceptually, opponents of sentimentalism must say what moral judgments are, such that they can be had dispassionately. What thought is a dispassionate person conveying, when she says, “Killing the innocent is morally bad?” Any attempt to give a reductive answer will be vulnerable to open-question worries. No descriptive substitute for the phrase “morally bad” leaves us with a sentence that is conceptually synonymous with the third.

Arguably, the open-question argument does not threaten sentimentalism. Let’s distinguish two kinds of open questions. First, given a certain attitude towards killing, one can still wonder whether killing really is morally bad. Second, given a certain attitude toward killing, one can wonder whether one is thereby regard-

ing it as morally bad. Reductive theories of value leave both questions open. If I form the attitude that killing cannot be willed as a universal law, I can still wonder both whether killing is bad and whether I am judging that it is bad. Sentimentalism leaves the first question open, but not the second. When experiencing outrage at killing, it seems impossible to wonder I am regarding killing as bad. I can of course wonder whether killing really is as bad as it seems. Such doubts can arise because I may not know the true source of the emotion I am feeling. Perhaps my outrage comes from some extraneous source (such as a bitter beverage), for example. But this open question does not threaten the thesis that moral judgments are constituted by sentiments. The only open question that poses such a threat would be one about what my attitude is, not one about whether my attitude is true. The fact that some sentiments are experienced as condemnatory effectively closes the question about whether someone experiencing those sentiments is adopting a moral stance. By analogy, imagine tasting a wine and wondering whether it really is delicious, while experiencing gustatory pleasure. We can have this thought (a thought about truth), because we can’t be sure where the pleasure came from (was it the wine or the company?). But we can’t experience gustatory pleasure and wonder whether we are, at that moment, finding the experience delicious. Thus, gustatory pleasure is plausible a component of deliciousness judgments.

The foregoing may look like a conceptual argument for sentimentalism. But it can also be construed as an empirical claim. The argument hangs on the premise that people experiencing outrage take themselves to be making moral judgments. This can be empirically tested. Indeed, all the evidence about people consulting their emotions when making moral judgments stands as evidential support. Merely making someone mad results in more negative moral attitudes. This can be interpreted as showing that, when people are angry, there is no question for them about whether they are holding something in negative moral regard. Conversely, it would be easy to show that people do not ne-

cessarily draw this inference when they form the judgment that something cannot be willed as a universal law. Opponents of sentimentalism owe us a positive account of evaluative thoughts that avoids open-question worries as successfully as sentimentalist accounts.

Opponents of sentimentalism might try to bypass this demand by offering a non-reductive account of moral judgments, treating thin moral concepts as primitives. That possibility, which was attractive to Moore, looks unmotivated given the empirical evidence for an emotional foundation. Every study suggests that emotions arise when we make moral judgments. All evidence also suggests that when emotions are eliminated, judgments subside as well. This does not prove that we can make moral judgments without emotions, but, by induction, it provides evidence. Some have argued that extant evidence is ambiguous about whether emotions are essential components of moral judgments or mere accompaniments, but I have suggested here that the former may provide a better explanation (and certainly better predictions) of the total pattern of data (Huebner et al. 2009; Waldmann et al. 2012). Until opponents of sentimentalism can identify some clear cases of moral judgments without emotions, they will be on the losing side of the debate. At the moment, there is no empirical evidence that this ever happens.

Notice too, that it would be relatively uninteresting to show that, under as-yet-unidentified and highly unusual conditions, people can make what look like moral judgments in the absence of emotions. The sentimentalist will reply that the vast majority of ordinary moral judgments are emotionally based. If moral vocabulary is occasionally used dispassionately, sentimentalists can ask whether the thoughts expressed on such occasions are of the same kind that we find, in study after study, in the usual cases. Upon finding a class of dispassionate judgments, one might do best to posit an ambiguity in the category. The sentimentalist can content herself with the project of providing a metaethics for garden-variety moral judgments, while leaving open the possibility that there may be psychological exotica, which conform to

the theories of their opponents. At the moment, there is no empirical evidence for such exotica.

More modestly, the empirically-minded sentimentalist might welcome an attempt to find evidence for opposing views. Little effort has been put into this task, though empirical claims for emotion-free moralizing are occasionally advanced. The most publicized example is Koenigs et al.'s (2007) study, which shows intact consequentialist judgments in patients who suffer from ventromedial prefrontal brain injuries, which are thought to impair emotion. But this description is misleading. As the authors note, ventromedial patients are highly emotional, and their most notorious symptom is that they are insensitive to costs when seeking rewards. Presumably, reward-seeking is an affectively grounded behavior. The fact that these patients make normal consequentialist judgments does not entail that they rely on reason alone, but rather on their positive emotions. Since these emotions cannot be easily regulated by negative feedback in ventromedial patients, they tend to be more consequentialist than healthy populations—that is, they are more willing to push a heavy man in a trolley's path in order to save five.

Will better empirical evidence for rationalism or externalist moral realism be forthcoming? I doubt it. Rationalists hold that we can arrive at moral judgments through reasoning. Unlike some sentimentalists, I think reasoning is important to morality. It is likely that we use reasoning to extrapolate from basic values to novel cases. But it is unlikely that we could use reasoning to derive basic moral values. Philosophers have tried to do this for centuries with no consensus behind any view. This might be described as a strong empirical argument by induction: thousands of smart, trained moral experts have failed to identify a line of reasoning that is widely recognized as providing adequate rational support for basic moral propositions. Moreover, when moral debates arise, there is little evidence that reasoning is efficacious on its own. Instead, societal transformations in values seem to arrive with political upheavals, economic revolutions, and generational change. Attitudes towards slavery changed with the indus-

trial revolution, women's suffrage came with a world war, and increase in support for gay rights correlates with the dissolution of traditional social roles and economic transformations that have made procreation more costly than abstinence. I don't mean to imply that there are no rational arguments for these liberation movements. Rather, I am suggesting that those arguments take hold only when social conditions are right. It is noteworthy, for example, that scientific racism appeared very late in the history of slavery, suggesting that slavery was not simply based on false beliefs about racial inequality. In fact many societies have enslaved their own people, and many proponents of scientific racism have been against slavery. Rather, advocacy of slavery seems to reflect a set of basic moral values that changed in recent history: values that say social standing can be determined by the lottery of birth. With industrialization, models of labor based on the idea of self-determination took hold, and the idea that birth should determine social standing began to wane. Of course, it hasn't disappeared altogether, but it has been tempered by the emergence of a new norm. Before industrialization, the idea that human beings are born equal and free might have seemed manifestly false, and thus it could have played no effective role in any argument against slavery. With industrialization, this premise gained appeal, and became the foundation of compelling arguments. Arguments are not inert, but they are only as good as the premises on which they are based, and the plausibility of those premises may depend on factors other than reasoning. It is possible that reasons have little role in driving basic values. And if so, then the recent broadening moral umbrella is not the result of a rational inference to the conclusion that our basic values cover more cases than we thought, but rather an irrational shift in basic values.

A realist might concede that such considerations threaten rationalism, but vie instead for a kind of intuitionist perspective, according to which basic moral truths are simply obvious. To me, this looks like a magical moral epistemology—one wonders what moral facts could be such that our moral sense could simply lock on to

them. It is also open to a damaging empirical objection. Phenomenologically, it is true that moral intuitions often seem immediate and unbidden, but this can be readily explained on a sentimentalist account. Emotions are conditioned (by training or evolution) to arise automatically and often intensely when certain actions, such as torturing babies, are considered. This gives an impression of immediacy without postulating any special contact with moral reality. Moreover, these intuitions vary from group to group. For example, there is empirical evidence that liberals and conservatives have divergent basic values (Graham et al. 2009). The presence of such foundational intuitions can be explained demographically, and their lack of convergence casts doubt on the existence of a moral faculty that reveals universal moral truths. In other words, intuitionism is vulnerable to a debunking argument. Social science coupled with sentimentalism provides a good explanation of deeply-held intuitions, so there is no need to suppose that these intuitions reflect anything deeper.

This point about moral variation, to which I will return, also counts against some forms of externalist moral realism. Advocates of that position sometimes suggest that objective moral facts can be established by identifying the external factors that best explain human moral behavior or judgments. If moral behavior and judgments vary from group to group, however, it is unlikely that we will find an external common denominator underlying these practices. Such a search also seems unnecessary given that we already have good explanations of moral behavior and judgments in terms of socially-conditioned sentiments.

None of these arguments are the nail in the coffin for externalist realist or rationalist theories. They merely illustrate the relevance of empirical results. The findings mentioned here must be explained. It is my contention that sentimentalism provides the best explanation of the findings I have reviewed, but further arguments and evidence could tip the balance in another direction.

2.2 Cognitivism vs. non-cognitivism

Let's move on to the second question on the metaethics decision tree: Are moral judgments truth-apt? As positioned on the tree, this is a

question that arises for sentimentalists, raised pressingly by the conclusion that moral judgments have a basis in the emotions. It is that conclusion that seems to put truth-aptness in jeopardy, since emotions have not traditionally been regarded as having accuracy conditions in the way regarded as allowing for truth. But, it should be noted that the question of truth-aptness could also be raised independently of sentimentalism. There are non-sentimentalist theories that deny truth-aptness (for example, one might say that moral judgments are imperative, while denying that they need be passionate), and there are non-sentimentalist theories that accept truth-aptness (the vast majority fall in this category). To keep things as neutral as possible, I will begin by asking whether there is any empirical evidence that moral judgments are non-cognitive, whether or not they are affect-laden.

The posing of this question is itself a degree of philosophical progress, because non-cognitivists too rarely reflect on the predictions of their view. Indeed, the most obvious empirical prediction fails resoundingly. If moral judgments do not aim at truth, we might expect them to have a non-declarative syntactic. For example, we might expect them to take the form of imperatives or exclamations. But they do not. In every language that I know of, moral judgments are expressed using declarative sentences, which should stand as a profound embarrassment to the theory. Granted, non-cognitivists sometimes propose elaborate logics to accommodate this fact, but it is surprising that they should have to do so. One would expect the surface grammar to reflect the non-cognitive form.

To push things further, one might look for more subtle linguistic evidence in favor of non-cognitivism. For example, some non-cognitivists assume that moral utterances have the illocutionary force of directives, such as orders, requests, or demands. Directives often occur in speech contexts that contain words that play a role in persuasion, such as “come!”, “let’s”, or “we encourage you...” To empirically test this kind of non-cognitivism, Olasov (2011) ingeniously used this technique for sociolinguistics, called corpus analysis. He used a set of such linguistic elements

that correlate with directive speech, such as those just mentioned, and he searched corpora of spoken and written texts for co-variance between these elements and moral terms. He calls the directive elements “suasion markers,” and the correlations between these and other linguistic items a “suasion score.” Non-cognitivism seems to predict a high suasion score, given the postulated directive function of moral judgments. This prediction fails. Not only is there no positive correlation between moral vocabulary and suasion markers, there is actually a negative correlation, which approaches significance. This negative relationship was observed in seventeen out of nineteen different categories of text that he examined. These results are preliminary—a first foray into empirical ethics—but they provide compelling evidence that moral discourse is not directive in nature.

Non-cognitivism entails that moral discourse does not aim to refer to facts in the world. This carries another linguistic prediction that can be readily tested. Certainly adverbs are used to indicate a focus on how things are in the world. These include “really,” “truly,” and “actually.” These words have other uses (“really” can be a term of emphasis), but they often play a role in emphasizing the factive nature of the modified phrase. Therefore, if non-cognitivism were true one might expect these words to rarely be used as modifiers for moral terms. To test this, I used Google search engine to search for and note the frequency of three phrases: “really immoral,” “truly immoral,” and “actually immoral.” To do this, I needed a baseline, and chose to compare “immoral” to a word widely believed to designate a objective feature of the world. I chose “triangular,” a classic primary quality, on a Lockean scheme. The results are as follows (as of March, 2013):

“really triangular”: 6,500 hits
 “really immoral”: 10,600 hits
 “truly triangular”: 4,920 hits
 “truly immoral”: 32,000 hits
 “actually triangular”: 21,600 hits
 “actually immoral”: 61,600 hits

Clearly, the adverbs that indicate a real-world focus are used more frequently for moral terms

than for terms designating objective physical features—over six times as common in the case of “truly.” I also tried the phrases “in truth,” “truthfully,” and “in actual fact”:

“truthfully triangular”: 6 hits
 “truthfully immoral”: 44 hits
 “in truth triangular”: 46 hits
 “in truth immoral”: 1,350 hits
 “in actual fact triangular”: 2 hits
 “in actual fact immoral”: 133 hits

These truth-tracking phrases modify “immoral” between 7 and 166 times more frequently than they modify “triangular.” Moreover, these differentials are misleadingly small because the base rate for “immoral” is far lower than “triangular” (6,910,000 hits as compared to 11,600,000). This was just an exploratory study, but there is a simple implication. Non-cognitivism makes linguistic predictions, and when those are tested, they do not seem to pan out. Non-cognitivists owe us evidence, or they must deny that their theory makes predictions, in which case it would cease to be falsifiable.

In response, non-cognitivists might claim that there is one crucial line of evidence in favor of the view, and it’s a line of evidence that we have already seen. In the previous section, I surveyed studies suggesting that morality is affect-laden. At the start of this section, I said the non-cognitivism is orthogonal to affect-ladenness, but some non-cognitivists would vehemently disagree with this. They would say that non-cognitivism *follows from* affect-ladenness. Emotions are traditionally regarded as feelings, and feelings are not traditionally believed to be representations of anything. If the thought that killing innocents is wrong is really a bad feeling about killing, then why think this thought has any truth conditions? Does a feeling of indigestion or irritation really refer?

This move might have been compelling in the early part of the twentieth century, but the last fifty years of emotion research have emphasized the intentionality of affect. Some philosophers have adopted cognitive theories of the emotions, according to which emotions are identical to judgments. Elsewhere I have argued

against such theories, in favor of the view that emotions are bodily feelings (Prinz 2004), but contemporary feeling theorists still insist that emotions aim to refer. Feeling sad, for example, can be understood as a downtrodden bodily state that represents loss. To say that the feeling represents loss is to say that it has the function of arising in response to losses, and hence carries the information that there has been a loss to a person who experiences it. In a like manner, pain may indicate tissue damage and fatigue may indicate energy depletion, even though pain and fatigue are bodily feelings. None of these feelings are arbitrary. They prepare an organism to cope with specific conditions or events. Emotions qua feelings are in the business of keeping us abreast about how we are faring. Each emotion has a different significance, and any one of them can misfire. I might be sad when there is no loss, or frightened when there is no threat. Such emotions would qualify as erroneous.

If emotions are in the business of representing, then there is no difficulty supposing that moral judgments are truth-apt. When we sincerely assert that, “Killing innocents is bad,” we express a negative feeling towards killing, and that feeling functions as a kind of visceral predicate. It attributes a property to killing (I will have more to say about this property below). In this sense, moral discourse may be much like other forms of emotional discourse. If we say that some food is icky, we express a feeling, while also attributing a property. For example, the feeling of ickiness might represent the property of noxiousness, or perhaps something more subjective, such as the property of causing nausea in the speaker. Someone who calls something “icky” need not know what property that feeling represents, but most language users probably recognize that in using this term we are attempting to say something about whatever it is that elicits the feeling. By analogy to “icky,” moral assertions can be understood as both expressive and predicative. It is a mistake, based on overly simplistic theories of emotions, to assume that feelings cannot play a semantic function. Once we see that feelings can represent properties and function as predicates,

non-cognitivism no longer looks like a serious option.

2.3 Realism vs. the error theory

It is one thing to say that moral assertions aim to represent and quite another to say that they succeed in doing so. It is possible that when we say that an action is immoral, we aim to ascribe a property to it, but we do not succeed in doing so. This is precisely what defenders of the error theory have claimed. So, even if the forgoing case for cognitivism succeeds, we must now descend the decision tree and ask whether moral judgments are ever true.

The error theory, which states that moral judgments are truth-apt but always false, was first promulgated by [J. L. Mackie \(1977\)](#). Mackie's argument begins with the premise that moral predicates aim to represent properties with two important features. The first is objectivity: moral properties are supposed to be the kinds of things that can obtain independent of our beliefs, desires, inclinations, and preferences. The second is action-guidingness: moral properties are supposed to be the kinds of things that compel us to act when we recognize them. Mackie's second premise is that these two features are difficult to reconcile. Objective properties are usually the kinds of things about which we can be indifferent. Mackie uses the term "queer" to describe properties that are both objective and action-guiding, and he also suggests that such queer properties would require an odd epistemology. For these reasons, he thinks we shouldn't postulate objective action-guiding properties. But, Mackie thinks that moral concepts commit to the existence of such properties, and, thus, that moral judgments posit properties that don't exist. Therefore, moral judgments are systematically false.

In recent years, the error theory has become popular among evolutionary ethicists ([Ruse 1991](#); [Joyce 2006](#)). Mackie's theory leaves us with a puzzle. Why do people make moral judgments if they are incoherent? Evolutionary ethicists purport to have an answer. They say that morality is an illusion that has been naturally selected because it confers a survival ad-

vantage. For example, if we believe that cheating others is objectively bad and that belief is action-guiding, then we will hold others accountable when they cheat, and we will resist cheating even when it might seem advantageous to do so. This reduces the likelihood of free riders and leads to an evolutionarily stable strategy—one that can foster cooperation and collective works. Evolutionary ethicists also typically endorse sentimentalism, suggesting that moral emotions have evolved to motivate such things as punishment and altruism. Mackie himself is not explicit about the role of emotions in his view, which makes it unclear what he means when he says that we perceive the discovery of alleged moral facts to be action-guiding. The link between judgments and emotions, emphasized by evolutionists, provides one answer.

The evolutionary addendum to Mackie's argument may look like an empirical reason for siding with the error theory. Natural selection is a well-confirmed process, emotions have some basis in evolution, and evolutionary models confirm that emotionally-grounded moral instincts would be adaptive. But there are empirical reasons for doubting the evolutionary story, and for doubting the key premises in Mackie's argument. Consequently, I think the case for the error theory fails.

The evidence for an evolved moral sense is underwhelming. A thorough critique cannot be undertaken here, but let me offer two broad reasons for doubt (for more discussion, see [Prinz 2007a](#)). First, there is little evidence for a moral sense in closely related species. Recall that moral judgments are underwritten by emotions such as anger, disgust, guilt, and shame. There is no evidence that the last three of these emotions exist in chimpanzees, and the anger they exhibit might better be described as reactive aggression, because there is little reason to believe chimps form robust tendencies to be angry about third party offences when they are not directly involved. Evolutionists point out that chimps engage in reciprocal altruism, and other forms of prosocial behavior, but these behaviors may not depend on any moral judgments. Indeed, psychopaths engage in reciprocal altruism ([Widom 1976](#)), and chimps often be-

have in ways that seem psychopathic; they can be extremely violent (Wrangham 2004) and indifferent to each others welfare (Silk et al. 2005).

Evolutionary ethicists might concede this and argue that morality evolved in the human species after we split from other primates. But this position is vulnerable to a second objection: there is good reason to think that morality in humans is learned. Moral judgments derive from emotions that originate outside the moral domain, such as disgust, which is first applied to noxious agents and later expended to the social domain, through conditioning (Prinz 2007a). Even guilt and shame may be learned byproducts of non-moral emotions: shame is related to embarrassment and guilt may be a blend of sadness and anxiety brought on by violating a social norm (Prinz 2005). These emotions and their range of application depend on extensive conditioning in childhood. Moral variation across cultures is considerable, as we will see, and shared moral values can be attributed to widespread constraints on building a stable society (for example, stable societies must prohibit wanton murder within the in-group). Moreover, there is no poverty-of-the-stimulus argument for morality; children receive ample “negative data” in the form of punishment, and they directly imitate values in their communities. As I argue in greater detail elsewhere, arguments for innate moral norms have been unconvincing (Prinz 2007a). This suggests that morality is learned, not evolved.

If morality is acquired through learning, then one cannot bolster Mackie’s argument by assuming that morality is the product of evolution. This alone does not undermine the error theory, however. Error theorists might abandon the evolutionary approach and try to explain systematic error by appeal to a learning story. There is some evidence that people tend to treat certain rules as universally binding, regardless of operative conventions. When asked whether it would be okay to hit a classmate if the teacher granted permission, children tend to say “no.” Turiel (1983, Ch. 7) who made this discovery, denies that such objectivist leanings are innate. Rather, he thinks children learn to

distinguish moral and conventional rules. Some subsequent authors have argued that the learning in question involves emotional conditionism (Blair 1995; Nichols 2004). Moral rules are acquired through the inculcation of emotions such as anger, guilt, and shame. There are strong negative feelings associated with hitting that don’t disappear when children imagine the teacher saying it is okay to hit. Violating social conventions may lead to other emotions, such as embarrassment, but these are mitigated when we move from one social setting to another. For example, wearing a hat at the dinner table might be frowned on in some circumstances, but not when wearing a birthday hat at a birthday party. The idea that moral rules are learned by emotional conditioning could also explain their motivational impact; emotions impel us to act, so emotionally grounded rules seem to carry practical demands. This analysis would explain both features emphasized by Mackie—action-guidingness and objectivity—without assuming that moral rules actually are objective. Thus, the error theory could get off the ground without assuming that morality is a product of evolution.

On closer scrutiny, however, this argument is not strong enough to rescue the error theory. It conflates objectivity with authority independence. It is true that children think hitting is wrong even when it is permitted, but that does not mean they think moral truths exist independently of subjective responses. Many of our subjective responses seem independent of what authorities happen to say—our preferences for food and music, for example. But we don’t necessarily infer that these things are objective. So it is a further empirical question whether objectivity is an essential feature of how we understand moral properties.

This brings us to the heart of Mackie’s argument. Should we grant his first premise that moral assertions entail objectivity? Empirically, the answer is a bit messy. When polled, many people assume that morality is objective, but many reject this assumption (Nichols 2004; Goodwin & Darley 2008). In survey studies, there is a nearly even split between objectivists and their opponents. Strikingly, belief in ob-

jectivity correlates with religiosity. Goodwin and Darley report that religious beliefs were the strongest predictor of objectivity that they were able to find. This suggests that belief in objectivity is not an essential part of moral competence, but is, rather, an explicitly learned add-on that most often comes with religious education. The authors also found that belief in objectivity goes down in cases of moral issues about which there is considerable public debate, such as abortion. This might be interpreted as showing, again, that objectivity is not a conceptual truth about the moral domain, but rather a negotiable add on, which can be abandoned in light of counter-evidence. Faith in objectivity goes up with certain religious beliefs (e.g., divine command theory), and goes down when confronted with the fact that decent, intelligent people have very different moral convictions. In Quine's terms, moral objectivism, when it is found, may be collateral information rather than an analytic truth—a belief about morality that we are willing to revise.

To test this hypothesis, I conducted a survey study in which I compared a moral predicate (*immoral*) to two natural kind terms (*beetle* and *tuberculosis*), which paradigmatically aim to designate objective properties, and to two terms that are often said to represent secondary qualities (*red* and *humorous*). If natural kind terms have a presumption of objectivity, then any threat to that presumption should lead people to conclude that those terms don't refer. Things are a little trickier with terms such as *red* and *humorous*: many people believe that they designate objective properties, but are willing to give up this assumption when presented with countervailing evidence. When told that there is no unifying essence to humor, people do not conclude that nothing is funny; they conclude that humorousness is a property that depends on our responses. In other words, objectivity is not analytically entailed by *humorous* or *red*. It is collateral information. My study was designed to see if *immoral* followed this same pattern.

A group of college undergraduates read the following vignette for the *immoral* case, with comparable vignettes for the other terms:

Suppose scientists discover that there are two kinds of things that people call immoral. Would it be better to say: (a) The term “immoral” is misleading, and it might be better to replace it with two terms corresponding to the two kinds of cases.

Or

(b) The fact that there are different cases is interesting, but doesn't affect the word. The fact that we react the same way to these two things is sufficient for saying they are both members of the same category; they are both immoral.

When given these options, 75% chose option (b) for *immoral*, resisting the first option which is tantamount to an error theory. Exactly as many chose option (b) for *red*, and a few more picked (b) for *humorous* (90%). In contrast, (a) was the dominant answer for the natural kind terms, *tuberculosis* and *beetles* (55% and 65% respectively). This suggests that people do not treat moral terms the way that they treat natural kind terms. Even if many people happen to think that morality is objective (as the studies by Nichols 2004, and Goodwin & Darley 2008, suggest), they are willing to give up on this belief without abandoning their moral concepts. They are willing to treat those concepts as response-dependent.

I think these results can be best interpreted as follows. Moral concepts are neutral about moral objectivity. People can acquire these concepts without any beliefs about what kinds of properties they designate. This neutrality begets a kind of resistance to error. If there are no objective moral properties, then it wouldn't follow that moral judgments fail to refer; it would mean only that they refer to response-dependent properties. Thus, it is all but guaranteed that some moral judgments will come out true, and to this extent the evidence favors moral realism (defined as the view that there are truthmakers for some moral judgments). Mackie mistakes a popular but dispensable belief about morality for an analytic truth. His error theory rests on an error. In fact, his argument for the error theory may rest on two

mistakes, the second of which we will come to presently. Of course, this is just one study, and other interpretations may be available, but it provides some evidence against Mackie's conceptual claim and shows how empirical findings might be used to explore whether moralizers are, as he suggests, committed to objectivism. Extant empirical evidence suggests otherwise.

2.4 Sensibility vs. moral sense

The survey study just described suggests that one can possess moral concepts without knowing whether moral judgments refer to properties that are objective. The survey also brings out the possibility that people are willing to accept the conclusion that moral truth depends on our responses. But the survey does not settle whether a response-dependent theory is true. This is the next question on the decision tree. As we have seen, Mackie thinks action-guidingness and objectivity are incompatible. This may suggest that he sees no room for a theory that combines moral objectivity with the view that moral judgments have motivational pull. This, however, is Mackie's second mistake. The hypothesis that morality has an emotional basis reveals a way out of Mackie's argument for incompatibility. Emotions are action-guiding in that they motivate us to act. But some emotions may also represent objective features of the world. Fear, for example, may represent danger, and danger may be an objective property. Emotions can represent objective properties in a motivating way: they simultaneously pick up on information while compelling us to respond adaptively. The fact that fear is action-guiding does not rule out the possibility that it is designed by evolution to track objective threats. Likewise, disgust is action-guiding but it may register real sources of contamination.

This brings us back to "icky." This emotionally-expressive term may refer to something objective, like contamination, or to something subjective, such as the tendency to cause feelings of nausea. We can ask whether ickiness is objective or subjective, even if we grant that the word "icky" is expressive. Expressive terms can have objective referents. Likewise, we can ask

this question about moral terms. This question frames a historical debate between Francis Hutcheson, who may have believed that our moral sentiments track objective moral truths, and David Hume, who suggests that morality depends on human responses. The claim that moral judgments track objective properties is called the moral sense theory. It seems to have been defended by Francis Hutcheson in the eighteenth century. It may even have been Kant's considered view, since he had an objective procedure for arriving at moral truth, but also insisted that every moral judgment is associated with a moral feeling. The moral sense view finds an analogue in contemporary authors who combine external standards of moral truth with motivationally charged moral psychologies (e.g., [Campbell 2007](#); [Copp 2001](#); see also [Railton 2009](#), who makes a modest move in that direction). The alternative view, which says that moral judgments refer to response-dependent properties, has been called the sensibility theory ([McDowell 1985](#); [Wiggins 1987](#)). We can now ask whether there is any way to decide between these options empirically.

I think there is some reason to favor sensibility over moral sense. For the moral sense theory to be true, there would have to be a candidate objective property to which our moral concepts could refer. Unfortunately, I cannot undertake a review of modern moral sense theories here, but I will offer, instead, a more general line of empirically-informed resistance. Moral rules are emotionally conditioned, and communities condition people to avoid a wide range of different behaviors. Within a given society, the range of things that we learn to condemn is remarkably varied. Examples include physical harm, theft, unfair distributions, neglect, disrespect, selfishness, self-destruction, insults, harassment, privacy invasions, indecent exposure, and sex with the wrong partners (children, animals, relatives, people who are married to other people). One might think that all of these wrongs have a common underlying essence. For example, one might propose that each involves a form of harm. But this is simply not true. Empirical evidence shows that people condemn actions that have no victims, such as

consensual sex between adult siblings and eating the bodies of people who die in accidents (Murphy et al. 2000). Furthermore, harm itself is a subjective construct. It cannot be reduced to something like physical injury. Privacy violations are regarded as a kind of harm, even though they don't hurt or threaten health, whereas manual labor is not considered a harm, but it threatens the body more than, say, theft. Similar problems arise if we try to define moral wrongs in terms of autonomy violations. Mandatory education violates autonomy, but it is considered good, and consensual incest is an expression of autonomy, but is considered bad.

Realists would no doubt resist some of these claims, but theirs is an uphill battle. On the face of it, morality lacks a common denominator. Empirical surveys of human values suggest that moral rules are a potpourri, which can be extended and contracted in any number of ways, with no fixed ingredients. Or rather, the common denominator is not a property shared by the things we condemn, but rather by the condemning itself. Moral sense theorists liken morality to perception, and, in so doing, they imply that there is an external feature of the world that our moral sentiments pick up on. But there is little reason to believe this. Unlike perception, there is massive variation in what we moralize, and there is a perfectly good explanation for this: the content of morality is determined by social conditioning rather than by the mind-independent world. Morality is not something we get by simply observing.

The foregoing is offered as an empirical challenge to moral sense theories, not a decisive refutation. Too often philosophers stick with examples of moral norms that clearly concern harm or violations of autonomy. This inflates optimism about a unifying essence. If one uses empirical methods to discover the full range of things that people actually moralize (such as victimless harms), the task of finding a unified essence looks much harder. Moral sense theorists might reply that this diversity is illusory. They might say, for example, that people would stop condemning victimless crimes on reflection. That claim is amenable to empirical testing, and so far the tests provide little support. For

example, Murphy et al. (2000) presented people with cases of incest and cannibalism where it was extremely salient that no one was harmed. They invited people to revise knee jerk moral intuitions and rule that, on reflection, these victimless actions are permissible. A piddling 20% revised accordingly, but 80% stuck to their original view. Moral sense theories seem to place their bets on the 20%. The challenge is to explain why the stubborn and considered opinions of the majority are performance errors of some kind.

Given the diversity of things about which people moralize, I think the sensibility theory is more promising than the moral sense theory. Wrongness is projected, not perceived. The property of being wrong is the property of causing negative sentiments, not a response-independent property that those sentiments are designed to detect. This conclusion follows from an inference to the best explanation. Empirically it looks as if there is no common essence to the things that we find morally wrong—a finding that is difficult to explain on the moral sense model, but easy to explain on the assumption that wrongness is response dependent. By analogy, imagine that we catalogue the things that make people laugh, and find that they lack a shared essence. This would imply that laughter does not pick up on an objective property. The things that we find funny are unified by the very fact that we are amused by them. Likewise for the things we find immoral: disapprobation carves the moral landscape.

2.5 Relativism vs. ideal observers

I have just been arguing that moral truth is response-dependent. Moral judgments can be true, but their truth depends on our sentiments. Something is immoral if it causes anger, disgust, guilt, and shame in us. But now we can ask, who does “us” refer to here? Whose sentiments determine moral truth? This brings us to the final question in the metaethics decision tree. Can divergent responses have equal claim to truth?

Empirical evidence strongly suggests that moral sentiments vary, both within and across

cultures. Within a culture, the clearest divisions are between political orientations. Liberals and conservatives have interminable debates, even when they are exposed to the same science and education. Research suggests that these debates come down to fundamental differences in moral values. Conservatives are much more likely than liberals to emphasize purity, authority, and preservation of the in-group in justifying their moral norms (Haidt 2007). These things are foundational for conservatives and largely irrelevant to liberals.

Across cultures, differences are even greater. Everything that we condemn is accepted somewhere else (such as slavery and torture), and things that have been condemned by other cultures (such as women's suffrage) have been embraced by us. There are cultures whose moral outlooks are dominated by considerations that we tend to downplay in the post-industrial West (sanctity and honor, for example), and ideals that are central to our moral outlook appear to be modern inventions (rights and the idea of human equality).

Descriptively, then, people do not seem to have the same moral values, within or across cultures. There is divergence in our sentiments. Some of this divergence might diminish if we filtered out cases where people were reasoning badly or on poor evidence, but there is ample evidence that disagreements remain among people who reason carefully and draw on the same factual knowledge. Indeed, if we filter for good reasoning, divergence might increase rather than decrease: consider professional normative ethicists, who are experts at reasoning but nevertheless arrive at varied and novel moral perspectives that neither converge with each other nor with the communities to which they belong.

I think such descriptive moral relativism provides support for metaethical moral relativism. This would be a terrible inference on its own, as every metaethics textbook points out, but the inference gains plausibility if bolstered by a premise I argued for above: moral truth is dependent on our responses. If responses vary, even under favorable epistemic conditions, and responses determine truth, then the truth of a

moral judgment can vary depending on whose values are being expressed.

The ethical universalist can resist this conclusion by offering an antidote to moral variation. The most natural strategy would be to defend universality by developing an ideal observer theory, and to argue that, under ideal epistemic conditions (which might include external factors as well as being an epistemically ideal agent), judges would arrive at the same set of moral values. This strikes me as woefully unlikely. Once we grant that sentimentalism is true, and that our sentiments track response-dependent properties, it's not clear how to settle on which observer is ideal. Two people who have the same factual knowledge may have different sentiments as a result of differences in temperament (Lovett et al. 2012), reward sensitivity (Moore et al. 2011), gender (Fumagalli et al. 2010), class (Côté et al. 2013), and age (Truett 1993). Whose sentiments are right? Moreover, the standard traits associated with ideal observation may be problematic in the moral domain. Should we consult someone who is disinterested when we know, empirically, that distance from a situation can lead to moral indifference? Should we consult someone who has not been conditioned by a particular culture when we know that innate sentiments are unlikely to deliver moral attitudes? Should we consult someone who attends to every detail of a case, when we know that framing, vivid description, and concreteness can alter moral judgments? These problems strike me as insuperable. There are no clear criteria for ideal observation and no reason to believe that careful observers would converge.

In posing this challenge, I am inviting ideal observer theorists to look at empirical findings and propose epistemic standards that would overcome the sources of variation mentioned here. Some ideal observer theories try to be empirically responsive in this way. For example, Smith (1994) advances the hypothesis that ideal rational agents would converge, but he also realizes that some readers might be reluctant to share his optimistic outlook. To quell these doubts he makes three empirical observations (p. 188): there is considerable moral con-

vergence already (he cites the existence of thick concepts as evidence: we all think brutality is bad and honesty is good); there has been moral progress (he cites slavery, among other examples); and entrenched disagreements often reflect faulty rationality, such as religious beliefs. Here, I think further empirical scrutiny would weaken Smith's case. Divergence is rampant, and people disagree on the scope of thick concepts (is torture brutal? is espionage dishonest?). Cases of (what we consider to be) moral progress are, I've noted, often driven by economic upheavals and other irrational factors, with reasoning playing a post-hoc role. Finally, disagreements remain after bad reasoning and religiosity are controlled for; the examples mentioned, in formulating the challenge include things such as temperament and framing effects. I think empirical evidence provides little reason to expect that rational and informed observers would deliver consistent verdicts.

In light of such worries, universalists might abandon the ideal observer theory and offer instead a procedural approach to consensus, arguing that people would and should converge if they arrived at their sentiments in the right way. For example, many people might agree that it is good to arrive at decisions democratically, taking multiple sentiments into consideration, and we might sentimentally endorse the outcome of democratically-resolved moral disputes. Though I cannot make the case here, I suspect the problems with such a procedural approach outweigh its prospects. Democratic decision-making does not result in moral consensus; it can even polarize. When such procedures increase consensus it is often through power and prestige rather than sentimental convergence. Our faith in democratic procedures may also be an expression of moral relativism rather than a solution. Democratic procedures are an historical anomaly, which emerged in the modern period with the rise of capitalism, and they have often been used to oppress minorities and to impose the values of the many over the few. Perhaps such procedures are an improvement over totalitarian forms of decision-making, but they do not remedy relativism. Indeed, as societies move towards consensus-building pro-

cedures, they may actually promote variation, leading to an endless proliferation of values and an ever widening gulf between those who cherish diversity and those who reside in more traditional societies. From a social science perspective, the prospects for a universal morality look grim.

Once the case for relativism is established, the question arises: relative to what? Are moral judgments relative to value systems? Are those systems individuated at the scale of cultures and subcultures or do they vary across individuals? Little empirical work has been done to address this question, but let me end with a suggestion about how to proceed. When examining the semantics of natural kind terms, philosophers have sometimes appealed to a linguistic division of labor (Putnam 1975). We defer to experts and thereby license them to adjudicate the boundaries between natural kinds. Now we can ask, is there such a thing as moral expertise? Do we appeal implicitly or explicitly to moral experts? Would we change our moral judgments if the designated members of our community told us we were morally mistaken? We don't know the answers to such questions, because moral expertise has not been intensively studied. I suspect there will be considerable individual differences, with members of more traditional societies showing more willingness to defer. But I also suspect that deference in the moral domain will be less prevalent than for natural kinds; we are more inclined to take ourselves as having morally authoritative insight. What is most clear, however, is whether the scope of the relativity depends ultimately on how we use moral concepts and terms; and this is something that can be investigated empirically. Naturalizing relativism will require the marriage of cultural anthropology and sociolinguistics. From the armchair, it is tempting to think there is a single true morality; introspective reflection tends towards solipsism.

3 Conclusion

Throughout this discussion, we have worked our way down a metaethics decision tree. I have

made a case for a relativist cognitivist sentimentalist sensibility theory. Admittedly, each of my arguments is only a first pass, and much more could be said for and against these positions. Many of the empirical findings that I have described are preliminary. My main goal here is not to make a decisive case for any position in metaethics. Rather, I am pleading for the relevance of empirical methods in doing this traditionally philosophical work. Moral philosophy is undergoing a process of naturalization. This has been felt most strongly in normative theory (e.g., the debate about the status of character in virtue ethics) and moral psychology (e.g., questions about how deontological and consequentialist judgments are made). I hope to have shown that the empirical work also bears directly on metaethical questions—questions about what, if anything, is the source of moral truth.

Empirical work cannot replace philosophical toil. We need philosophy to pose questions and identify possible theories. Experimental design is itself a kind of philosophical reasoning, and it takes considerable argumentation to move from data to theory. Naturalization is supplementation, not usurpation. But it is not just supplementation. The empirical arsenal may just be our best hope for adjudicating philosophical debates. Reflection can delineate the logical space, but we need observation to locate ourselves therein. Philosophers have always relied on observation, in some sense, but scientific methods allow us to observe processes that are unconscious, inchoate, or distant in space and time. Empirical studies can test the content, prevalence, and malleability of intuitions, and they can also tell us where our intuitions come from—a question of central metaethical concern. We should embrace any tools that help us resolve the questions that we are employed to answer. A century ago, there was a linguistic turn, and philosophers began to treat traditional philosophical problems as amenable to semantic analysis. Around the same time, the boundary between philosophy and psychology was still blurred, and journals such as *Mind* published articles that we might now classify as psychological. Such crossovers

fell out of fashion, however, and it has taken a century to get back to this incipient moment. With the linguistic turn, Anglophone philosophers became convinced that we should all learn logic because it would help us make progress. Logic did help, and it did not undermine philosophy. Now, we can encourage all philosophers to learn about methods and results used in the relevant social and physical sciences. The payoff of this naturalistic turn may be vastly greater than the linguistic turn. Science, not formal logic, is positioned to tell us whether morality is a human construction.

Acknowledgments

This discussion has benefited immeasurably from the feedback of anonymous referees and from Ying-Tung Lin, Jessica McCormack, Thomas Metzinger, and Jennifer Windt. I am grateful for their close reading and helpful suggestions.

References

- Blackburn, S. (1998). *Ruling passions*. Oxford, UK: Oxford University Press.
- Blair, R. J. R. (1995). A cognitive developmental approach to morality: Investigating the psychopath. *Cognition*, 57 (1), 1-29.
- Borg, J., Hynes, C., van Horn, J., Grafton, S. & Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: An fMRI investigation. *Journal of Cognitive Neuroscience*, 18 (5), 803-817.
- Boyd, R. (1988). *Essays on moral realism*. Ithaca, NY: Cornell University Press.
- Brandt, R. (1959). *Ethical theory: The problems of normative and critical ethics*. Englewood Cliffs, NJ: Prentice-Hall.
- Campbell, R. (2007). What is moral judgment? *Journal of Philosophy*, 104 (7), 321-349.
- Copp, D. (2001). Realist-expressivism: A neglected option for moral realism. *Social Philosophy and Policy*, 18 (2), 1-43. [10.1017/S0265052500002880](https://doi.org/10.1017/S0265052500002880)
- Côté, S., Piff, P. K. & Willer, R. (2013). For whom do the ends justify the means? Social class and utilitarian moral judgment. *Journal Of Personality And Social Psychology*, 104 (3), 490-503. [10.1037/a0030931](https://doi.org/10.1037/a0030931)
- Doris, J. M. (1998). Persons, situations, and virtue ethics. *Noûs*, 32 (4), 504-530.
- Eskine, K. J. (2011). *From perceptual symbols to abstraction and back again: The bitter truth about morality*. New York, NY: Doctoral dissertation, Department of Psychology, City University of New York.
- Eskine, J. K., Kacinik, A. N. & Prinz, J. J. (2011). A bad taste in the mouth: Gustatory disgust influences moral judgment. *Psychological Science*, 22 (3), 295-299. [10.1177/0956797611398497](https://doi.org/10.1177/0956797611398497)
- Firth, R. (1952). Ethical absolutism and the ideal observer. *Philosophy and Phenomenological Research*, 12 (3), 317-345.
- Flanagan, O. (1991). *Varieties of moral personality: Ethics and psychological realism*. Cambridge, MA: Harvard University Press.
- Fumagalli, M. M., Ferrucci, R. R., Mameli, F. F., Marcegaglia, S. S., Mrakic-Sposta, S. S., Zago, S. S. & Priori, A. A. (2010). Gender-related differences in moral judgments. *Cognitive Processing*, 11 (3), 219-226. [10.1007/s10339-009-0335-2](https://doi.org/10.1007/s10339-009-0335-2)
- Gibbard, A. (1990). *Wise choices, apt feelings*. Cambridge, MA: Harvard University Press.
- Goodwin, G. P. & Darley, J. M. (2008). The psychology of meta-ethics: Exploring objectivism. *Cognition*, 106 (3), 1339-1366. [10.1016/j.cognition.2007.06.007](https://doi.org/10.1016/j.cognition.2007.06.007)
- Graham, J., Haidt, J. & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96 (5), 1029-1046.
- Greene, J. D. (2007). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.) *Moral psychology, Vol. 3: The neuroscience of morality, emotion, disease, and development*. Cambridge, MA: MIT Press.
- Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M. & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, 293 (5537), 2105-2108. [10.1126/science.1062872](https://doi.org/10.1126/science.1062872)
- Haidt, J. (2007). The new synthesis in moral psychology. *Science*, 316 (5827), 998-1002. [10.1126/science.1137651](https://doi.org/10.1126/science.1137651)
- Hare, R. D. (1993). *Without conscience: The disturbing world of the psychopaths among us*. New York, NY: Pocket Books.
- Huebner, B., Dwyer, S. & Hauser, M. (2009). The role of emotion in moral psychology. *Trends in Cognitive Sciences*, 13 (1), 1-6. [10.1016/j.tics.2008.09.006](https://doi.org/10.1016/j.tics.2008.09.006)
- Joyce, R. (2006). *The evolution of morality*. Cambridge, MA: MIT Press.
- Kant, I. (1797). *The metaphysics of morals*. M. J. Gregor (Trans.). Cambridge, UK: Cambridge University Press.
- Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M. & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgments. *Nature*, 446, 908-911. [10.1038/nature05631](https://doi.org/10.1038/nature05631)
- Kornblith, H. (Ed.) (1985). *Naturalizing epistemology*. Cambridge, MA: MIT Press.
- Kramer, P. D. (1993). *Listening to Prozac*. New York, NY: Viking.
- Lovett, B. J., Jordan, A. H. & Wiltermuth, S. S. (2012). Individual differences in the moralization of everyday life. *Ethics & Behavior*, 22 (4), 248-257. [10.1080/10508422.2012.659132](https://doi.org/10.1080/10508422.2012.659132)
- Mackie, J. L. (1977). *Ethics: Inventing right and wrong*. London, UK: Penguin.
- McDowell, J. (1985). *Morality and objectivity*. London, UK: Routledge & Kegan Paul.
- Mikhail, J. (2002). Aspects of the theory of moral cognition: Investigating intuitive knowledge of the prohibition of intentional battery and the principle of double effect. *Economics Research Paper*, 762385. <http://ssrn.com/abstract=762385>

- Moore, A. B., Stevens, J. & Conway, A. A. (2011). Individual differences in sensitivity to reward and punishment predict moral judgment. *Personality And Individual Differences*, 50, 621-625. [10.1016/j.paid.2010.12.006](https://doi.org/10.1016/j.paid.2010.12.006)
- Murphy, S., Haidt, J. & Björklund, F. (2000). Moral dumbfounding: When intuition finds no reason. Unpublished manuscript, Department of philosophy, University of Virginia.
- Nichols, S. (2004). *Sentimental rules: On the natural foundations of moral judgment*. New York, NY: Oxford University Press.
- (2008). Sentimentalism naturalized. In W. Sinnott-Armstrong (Ed.) *Moral psychology: The cognitive science of morality: Intuition and diversity* (pp. 255-274). Cambridge, MA: MIT Press.
- Olasov, I. (2011). Register variation and the moral cognitivism debate. Unpublished manuscript, City University of New York, Graduate Center.
- Prinz, J. J. (2004). *Gut reactions: A perceptual theory of emotion*. New York, NY: Oxford University Press.
- (2005). Imitation and moral development. In S. Hurley & N. Chater (Eds.) *Perspectives on imitation: From cognitive neuroscience to social science*. Cambridge, MA: MIT Press.
- (2007a). Is morality innate? In W. Sinnott-Armstrong (Ed.) *Moral psychology, vol 1: Evolution of morals* (pp. 367-406). Cambridge, MA: MIT Press.
- (2007b). *The emotional construction of morals*. Oxford, UK: Oxford University Press.
- Putnam, H. (1975). The meaning of “meaning”. *Mind, Language and Reality: Philosophical Papers, Volume 2* (pp. 215-271). Cambridge, UK: Cambridge University Press.
- Quine, W. V.O. (1969). *Ontological relativity and other essays*. New York, NY: Columbia University Press.
- Railton, P. (1993). What the non-cognitivist helps us to see the naturalist must help us to explain. In J. Haldane and C. Wright (Eds.), *Reality, Representation and Projection* (pp. 279-297). Oxford, UK: Oxford University Press.
- (2009). Internalism for externalists. *Philosophical Issues*, 19 (1), 166-181. [10.1111/j.1533-6077.2009.00165.x](https://doi.org/10.1111/j.1533-6077.2009.00165.x)
- Ruse, M. (1991). A companion to ethics. In P. Singer (Ed.) *A companion to ethics* (pp. 500-510). Oxford, UK: Blackwell.
- Sayre-McCord, G. (Ed.) (1988). *Essays on moral realism*. Ithaca, NY: Cornell University Press.
- Schmidt, E. Z. & Bonelli, R. M. (2008). Sexuality in Huntington’s disease. *Wiener Medizinische Wochenschrift*, 158 (3-4), 84-90. [10.1007/s10354-007-0477-8](https://doi.org/10.1007/s10354-007-0477-8).
- Seidel, A. & Prinz, J. J. (2013a). Sound morality: Irritating and icky noises amplify divergent moral domains. *Cognition*, 127 (1), 1-5. [10.1016/j.cognition.2012.11.004](https://doi.org/10.1016/j.cognition.2012.11.004)
- (2013b). Mad and glad: Musically induced emotions have divergent moral impact. *Motivation and Emotion*, 37 (3), 629-637. [10.1007/s11031-012-9320-7](https://doi.org/10.1007/s11031-012-9320-7)
- Silk, J. B., Brosnan, S. F., Vonk, J., Henrich, J., Povinelli, D. J., Richardson, A. F., Lambeth, S. P., Mascaro, J. & Schapiro, S. J. (2005). Chimpanzees are indifferent to the welfare of other group members. *Nature*, 435, 1357-1359. [10.1038/nature04243](https://doi.org/10.1038/nature04243)
- Smith, M. (1994). *The moral problem*. Oxford, UK: Blackwell.
- Truett, K. R. (1993). Age differences in conservatism. *Personality and Individual Differences*, 14 (3), 405-411. [10.1016/0191-8869\(93\)90309-Q](https://doi.org/10.1016/0191-8869(93)90309-Q)
- Turiel, E. (1983). *The development of social knowledge: Morality and convention*. Cambridge, UK: Cambridge University Press.
- Waldmann, M. R., Nagel, J. & Wiegmann, A. (2012). Moral judgment. In K. J. Holyoak & R. G. Morrison (Eds.) *The Oxford handbook of thinking and reasoning*. Oxford, UK: Oxford University Press.
- Wastell, C. & Booth, A. (2003). Machiavellianism: An alexithymic perspective. *Journal of Social and Clinical Psychology*, 22 (6), 730-744. [10.1521/jscp.22.6.730.22931](https://doi.org/10.1521/jscp.22.6.730.22931)
- Widom, C. S. (1976). Interpersonal conflict and cooperation in psychopaths. *Journal of Abnormal Psychology*, 85 (3), 330-334. [10.1037/0021-843X.85.3.330](https://doi.org/10.1037/0021-843X.85.3.330)
- Wiggins, D. (1987). A sensible subjectivism. In D. Wiggins (Ed.) *Needs, values, truth: Essays in the philosophy of value* (pp. 185-214). Oxford, UK: Blackwell.
- Woolfolk, R. L. (2013). Experimental philosophy: A methodological critique. *Metaphilosophy*, 44 (1-2), 79-87. [10.1111/meta.12016](https://doi.org/10.1111/meta.12016)
- Wrangham, R. (2004). Killer species. *Daedalus*, 133, 25-35.

Conceptualizing Metaethics

A Commentary on Prinz

Yann Wilhelm

In this commentary on Prinz's "Naturalizing Metaethics" I shall first look briefly at his methodological assumptions. I will argue that Prinz's approach is more radical and less conciliatory between analytical and empirical approaches than it seems from his own description. In the second part of my commentary, I shall look at one possible objection to Prinz's sentimentalism: the evidence he presents does not provide the needed modal strength for sentimentalism. I shall present two example of this objection, and argue that Prinz's own depiction doesn't adequately represent it. I shall then use the helpful distinction offered by Jon Tresan between *de dicto*- and *de re*-internalism to analyze underlying problems in the objection. I will present another way of reacting to it, which I think fits nicely with Prinz's naturalized methodology. In the last part, I shall look at his critique of non-cognitivism. Prinz argues that non-cognitivism makes certain linguistic predictions that turn out to be wrong: if non-cognitivism were true we would expect our moral language to reflect this. I will argue that there are many forms of non-cognitivism that predict this surface grammar. The key idea is that non-cognitivism entails a pragmatic theory of moral language. I then offer a speculative explanation about why the moral language has its surface form. This speculation, I argue, has at least the same amount of plausibility as cognitivist theories. Furthermore, this possible explanation is open to empirical investigation. I agree with Prinz that, ultimately, metaethical theories should be tested against empirical evidence. Prinz presents conceptual and empirical work as mutually enhancing enterprises. My commentary is, I hope, a small contribution highlighting the conceptual side of the coin.

Keywords

Cognitivism | De dicto-internalism | De re-internalism | Metaethics | Methodological naturalism | Motivational internalism | Non-cognitivism | Sentimentalism

Commentator

Yann Wilhelm

ywilhelm@students.uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Target Author

Jesse Prinz

jesse@subcortex.com
City University of New York
New York, NY, U.S.A.

Editors

Thomas Metzinger

metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Metaethics under empirical scrutiny

Prinz proposes to naturalize metaethics. Metaethics is traditionally regarded as a second-order discourse about ethics. Where normative ethics asks what is good and what is bad, what we should or shouldn't do, metaethics asks the question of what morality is itself (DeLapp 2011). Its subject is the ontology of moral properties, the semantics of moral discourse, the epistemic foundation of moral judgments and the psychology of moral opinions. These different aspects are highly interrelated—answers in one area influence questions asked in others.

There are many different ways to tackle the question of what morality itself actually is. Prinz

characterizes metaethics as being concerned with the foundations of moral judgments (Prinz this collection, p. 1). This is his starting point, which shapes his decision tree. He acknowledges that one could arrange the tree in different ways, depending on which aspect one wants to pull into focus.

Prinz's primary goal is to show that every question in the decision tree is empirically tractable (this collection, p. 1). This is his *methodological naturalism* (p. 2).¹ He argues that we

¹ He contrasts this with *metaphysical naturalism* and *semantic naturalism*. The former says that everything there is belongs to the natural world. The latter tries to reduce concepts from various domains in terms that are more likely to be naturalized in the metaphysical sense.

should study the domain of metaethics empirically. He wants to test “[...] theories derived from philosophical reflection against the tribunal of empirical evidence” (Prinz [this collection](#), p. 5).

Metaethics, according to him, is not the sole matter of armchair reflection. This seems natural when we characterize metaethics as the question of what morality itself is. But that goes against the view that metaethics—or philosophy in general—is not concerned with what actually is the case, but with what *must* be the case. What are the *necessary* conditions of morality? On this view, metaethics is concerned with statements that hold *a priori*. Most of the time this means deriving knowledge from reflection upon the meaning of our concepts. This method of *conceptual analysis* had been at the core of philosophy since the *analytic turn* (Prinz [this collection](#), p. 3).

Against this turn Prinz sets the *empirical turn* ([this collection](#), p. 3). He describes this development as an enrichment of the philosopher’s tool box. Where conceptual considerations help us to formulate theories and flesh out the differences between different views, empirical methods confirm the theories derived from this work. The former pose questions and formulates possible answers; the latter test those answers. Prinz emphasizes that empirical and traditional approaches are not opposed to one another ([this collection](#), p. 5). Rather, they complement each other. They’re more like opposing points on a continuum of methods for exploring the world.

It is important to see that this view is not as conciliatory between traditional analytic philosophy and empirical philosophy as it might seem. It does not leave room for *a priori* armchair reflection. In fact, Prinz even regards conceptual analysis as an empirical task: “[A]rmchair conceptual analysis can be characterized as an introspective memory retrieval process. As such, it can be regarded as a form of observation” (2008, p. 191).

When Prinz speaks of “traditional methods”, he does not include conceptual analysis as an *a priori* enterprise. Rather, he is referring to various tools, for example formal semantics or logic, which help us articulate theories. They are tools for exploring the natural world, from

which we gain knowledge only through experience. Prinz is a radical empiricist at heart.

An empirical scientist could ask: “What differentiates this from my own work?” For she, too, reflects upon different theories, how they relate to each other, formulates questions, and so on. This is an important part of scientific, empirical work. I think Prinz would agree. An important upshot of his naturalized philosophy is that there are no clear-cut borders between philosophy and psychology (Prinz 2008, pp. 204–206). They are different disciplines not because of their different subject areas or methods but for pragmatic reasons. They are different *academic* disciplines, shaped by sociological and historical processes. The borders between the different disciplines become blurred in the empirical turn. According to Prinz, this is a good thing.

I think this the real strength of Prinz’s approach. Arguably many disciplines are divided largely by pragmatic differences, like education and academic organization. Instead of demarcating different approaches, instead of drawing sharp lines between them, Prinz proposes that we unite them in the search for explanations of the natural world.

Prinz’s target article is a very good example of this approach. Here I want to make a few remarks in the spirit of Prinz’s own methodology. In the next section I will focus on a specific objection against Prinz’s answers to the first question in the decision tree. I think that it can clarify some consequences of his methodological naturalism for metaethics.

2 Internalism and modal strength

In this section I discuss Prinz’s answer to a potential objection to his sentimentalism, namely, that the evidence lacks *modal strength*. In fact, objections of this kind have already been raised against Prinz’s and other naturalistic metaethical theories already. I shall first argue that his answer doesn’t get to the heart of the objection. Second, I propose a way in which Prinz can and should answer it. To do this I shall present two instances where this objection has been made. A helpful distinction by Jon

Tresan will then show that there are actually two kinds of internalist theses at play here. Only one of these is really relevant for Prinz's naturalized metaethics, I shall argue. The objection then loses its force in light of Prinz's project of a naturalistic methodology. The following reasoning can also be seen as a small case study in recent (naturalized) metaethics.

The first question in Prinz's decision tree is whether moral judgments are essentially affect-laden or not. This is Prinz's take on the internalist-externalist debate.² This debate is a classical debate in metaethics that can be traced back to the British moralists (Darwall 1995). It concerns the question of whether *motivation* is *internal* or *external* to moral judgments. Do moral judgments necessarily involve motivation to act accordingly? Or does the motivation come from a desire external to them (e.g., the desire to be a good person)?³

Prinz advocates a position that he calls sentimentalism:

Sentimentalism =_{DF} Moral Judgments essentially involve affective states, such as emotions, in one of two ways: such states as constituent parts of moral judgments (traditional sentimentalism); or moral judgments are judgments about the appropriateness of such states (neo-sentimentalism). (Prinz this collection, p. 6)

The evidence for a link between moral judgments and emotions is overwhelming (Prinz this collection, p. 10). But is it enough to warrant a stronger relation than mere accompaniment? Even if we grant Prinz the interpretation that affective states are not only mere *consequences* of moral judgments, could we not still question whether they are essential components of moral judgments? The objection is this: the empirical evidence lacks *modal strength* to support senti-

mentalism. Even if all our ordinary moral judgments are based on emotions, it could still be *possible* to judge dispassionately (Prinz this collection, p. 13). Therefore the evidence doesn't support sentimentalism.

Prinz answers that the empirical evidence gives us enough reason to infer that we *cannot* make moral judgments without emotions: "Every study suggests that emotions arise when we make moral judgments. All evidence also suggests that when emotions are eliminated, judgments subside as well" (Prinz this collection, p. 13).

According to Prinz, the theory that emotions are essential components of moral judgments explains the total pattern of data better than its rivals (this collection, p. 14). Furthermore, he argues that the sentimentalist can accept psychologically exotic cases, in which the connection between moral judgments and emotions doesn't occur, which conform rival theories.

This answer, I argue, misses the real core of the objection. Prinz confronts it upfront and just states what it questions. He puts the objection in the following way:

The evidence shows that emotions are often consulted when making moral judgments, but this leaves open the possibility that we might also make moral judgments dispassionately under circumstances that have not yet been empirically explored. (Prinz this collection, p. 13)

But this does not represent the objection adequately. The objection doesn't rest on possible, not-yet-found empirical evidence against sentimentalism. Rather, it rests on opposing ideas about what kind of modal strength claims about the relation between moral judgments and emotions should possess. At the heart of this objection there is no disagreement about the empirical evidence, but an opposition in the underlying methodology.

Adina Roskies, for example, accepts that "[...] those [brain] areas involved in moral judgments normally send their output to areas involved in affect, resulting in motives

² Although he doesn't explicitly put it like this, I think it's safe to frame it in this way. The option that denies affect-ladenness is called "externalist moral realism", and he states in various places that emotions are motivating or action-guiding (Prinz this collection, pp. 8, 11, 21). And one answer to the third question is a position called "internal realism". What I say about internalism in the following therefore applies equally to Prinz's sentimentalism. See also Prinz (2006), where he explicitly states motivational internalism.

³ See Björklund et al. (2012) for a short overview.

that in some instances cause us to act” (2008, p. 192).

But she thinks that this is not enough for internalism to be true.⁴ In her view there is a connection between the cognitive and the affective system, but “this link is causal and thus contingent and not constitutive” (Roskies 2008, p. 192). In this sense the connection, according to her, is not necessary.

Antti Kauppinen sees the difference between internalism and externalism in a similar way. He depicts internalism as saying that there is a link between moral judgments and motivation that holds a priori and with conceptual necessity. externalism, in contrast, is the view that this link is contingent and a posteriori (Kauppinen 2008, p. 3). For Kauppinen, every internalist position then becomes an externalist position if it weakens the modality of the claim. When a metaethical account doesn’t claim that the connection between moral judgments and motivation holds a priori and by necessity, it is an externalist account. No amount of empirical data can refute this criticism.

In Kauppinen’s case the disagreement with Prinz about the underlying methodology is clear. He reacts to the proposal by Roskies, Prinz, and Alfred Mele (among others) that we clarify the debate empirically (Kauppinen 2008, 4). Because of his definitions of internalism and externalism as conceptual necessary claims he argues that “[...] findings in either actual or fictional experimental psychology or neuroscience have little relevance to the debate” (Kauppinen 2008, p. 4).

Kauppinen is opposed to methodological naturalism in philosophical moral psychology (2008, p. 4). That is why he would not be satisfied with Prinz’s answer to this objection. Against him, Prinz would have to defend his metaethical naturalism. Interestingly enough, Roskies, on the other hand, thinks that we *can* clarify metaethical debates empirically.

In what follows I shall show how I think Prinz should meet this objection. Furthermore,

⁴ Her critique is directed at internalism, not sentimentalism. But I regard both positions as similar enough to treat Roskies’s critique as an argument against Prinz’s sentimentalism (see also above). At the core of both positions is the connection between moral judgments and affective (motivational) states.

I will argue that everyone who wants to apply empirical data to metaethical debates, such as, e.g., Adina Roskies, should side with Prinz on his methodological naturalism and accept internalism as a true a posteriori theory about moral judgments.

I will now present an analysis of the internalism–externalism debate offered by Jon Tresan that I think will be very helpful here (2009). He distinguishes different formulations of internalism along various dimensions. He claims that a very important distinction has been overlooked: most philosophers in the debate neglected the difference between the modality of the internalist claim and the stated relation between moral opinions and motivation. According to Tresan, there are two different kinds of necessity that can occur in such claims: wide-scope necessity, which operates over the entire proposition—*de dicto*—and narrow-scope necessity, which operates over the predicate—*de re* (Tresan 2009, p. 54). The first operates on the dimension of *Modality* and the second on the dimension of *Relation* (Tresan 2009, p. 55).

For example, the statement that parents have children can be formulated with both kinds of necessities:

Necessarily, parents have children (*de dicto*).

Parents have, necessarily, children (*de re*).

In the first case the proposition that parents have children is stated as holding necessarily. Parents have children, otherwise they would not be called parents. If someone has a child, she is a parent. But the second statement says that people who are parents have their kids necessarily. But this is obviously false. John and Mary don’t have their children necessarily. They could easily never have had any children at all. True, they would not, then, be parents – but the fact that they are parents may have, initially, been quite accidental. We can easily see that there is a difference between *de dicto*- and *de re*-necessities because these two statements can have different truth-values at the same time.

With this distinction at hand we can distinguish two different internalist theses: a strong Modality/weak Relation or *de dicto*-internalism, and a weak Modality/strong Relation or *de re*-internalism. The former states that, with necessity, there is a connection between moral judgments and motivation. The latter says that there is a necessary connection between these two things.

Tresan uses this distinction to argue that something has gone fundamentally wrong in the internalism–externalism debate. The neglect of the two features has led to the *internalist fallacy*: the strength in Modality of an internalist claim was taken to be strength in Relation, which led to an overestimation of the epistemic value of the claim (Tresan 2009, p. 55). The classical debate stated the connection between moral judgments and motivation in terms of conceptual necessity (a *de dicto*-internalism) (see Roskies’s and Kauppinen’s accounts above). Arguments for this claim were supposed to evoke the intuition that no one can make a moral judgment without being motivated to act. If we have such intuitions, the arguments go, the connection is a conceptual necessity. Likewise, arguments against this internalist claim consisted in thought experiments that were supposed to evoke contrary intuitions.

From Tresan’s distinction follows that claims with *de dicto* necessity are claims about our concepts and not about the subject matter (2009, p. 57). *De dicto*-internalism, then, is a claim about our concept “moral judgment” and *de re*-internalism a claim about the subject matter—the phenomenon of moral judgments.

Returning to Prinz (and to Roskies’s proposal), we can now see that there are really two empirical questions we can ask: First, what is our concept of “moral judgment”? And second, what are moral judgments? Traditionally the first was not regarded as an empirical question. Philosophers probed their intuitions and just assumed that others shared them. Prinz, on the other hand, regards these kinds of questions as empirical in nature and presents his own survey studies that probes *folk intuitions*. He concludes that most people do consider emotions necessary for moral judgments (Prinz this collection,

p. 10; for other studies on this with different results see also Nichols 2002, p. 22; Strandberg & Björklund 2013, p. 325; Björnsson et al. 2014, p. 16).

These studies can answer the first question regarding our concept of moral judgments. But, as Prinz rightly points out, people could be wrong (Prinz this collection, p. 10). These studies do not tell us anything about the subject matter. This is a further point Tresan makes. He argues that even if we have internalist intuitions this is not enough to support internalism. He argues that strength in modality is not interesting for a substantial theory of moral opinions. A claim with strong modality doesn’t tell us more about the subject of the claim than the same claim without it. That, necessarily, bachelors are unmarried (*de dicto*) tells us nothing more than that they need to be unmarried to be called bachelors. It’s a claim about our concept “bachelor”. It tells us simply that the subjects are unmarried—the same as this exact claim without modality tells us. But if bachelors were necessarily unmarried (*de re*) this would be bad news for the subjects and would tell us something substantial about them—that they’re essentially unmarried, that they, the individuals, are unable to be married. He concludes that “[i]f we are interested in the nature of the Subject Matter, we must look to Relation not Modality” (Tresan 2009, p. 57; emphasis in original).

Only an internalist claim with a strong relation is interesting. But Tresan thinks that there are no arguments for a *de re*-internalism, which would tell us something interesting and substantial about the subject matter. A *de re*-internalism that states a strong Relation is wrong. This is because our intuitions regarding moral judgments and motivation can only support a *de dicto* internalism (Tresan 2009, p. 64). And traditional arguments for internalism provoke only such intuitions.

I think it is clear that Tresan misses one important possible source of evidence for a strong relation: empirical evidence. Here lies the connection to Prinz’s work. The empirical findings, which he collected, all point to a strong relation between moral judgment and affective states. I take Prinz to be looking for a strong

Relation when he says that emotions are an “essential component” of moral judgments ([Prinz this collection](#), p. 12).

What I have tried to show here is the following. Prinz raises a potential objection against his own sentimentalism: the relation between moral judgments and emotions lack modal strength. He answers by saying that we have enough evidence to conclude their necessary connection. I argued that this is not a satisfying answer because it misses the core of the objection.

I think the evidence that he has collected points to a strong Relation between moral judgments and affective (motivational) states. Therefore Prinz has an answer to objections that call this strong relation into question. But this is not an answer to an objection that operates with a *de dicto* internalism. Underlying these objections is an opposition to methodological naturalism in general. [Antti Kauppinen](#) is one example of someone holding this position (2008, p. 4). Kauppinen does not think we should ask what moral judgments *actually* are. In his view, metaethics is concerned with what moral judgments *necessarily* are. “This takes us from the realm of the actual to the realm of the metaphysical or conceptually possible, and thus beyond the empirical and the observable” ([Kauppinen 2008](#), p. 22).

The evidence that Prinz presents in the target paper doesn’t suffice to refute this position. But I hope to have shown that this need not be a cause of concern for Prinz, because this kind of necessity takes us away from the subject matter. At the heart of Prinz’s account lies an interest in moral judgments as a natural phenomenon that we should study by empirical means.

Adina Roskies, on the other hand, is sympathetic to empirical philosophy. One of her aims in the internalist–externalist debate was to show that “[...] moral philosophy need not be, and perhaps ought not be, exclusively a priori” ([Roskies 2003](#), p. 2003).

But this is in contrast to her understanding of the required modality of the internalist claim, as I tried to show using Tresan’s analysis. If we want to clarify those kinds of debates em-

pirically, it’s not enough to just take traditional philosophical claims and look for evidence in their favor or evidence that can refute them. We have to formulate them as a posteriori synthetic claims that are part of a bigger explanatory project ([Björnsson 2002](#), p. 329).

I hope that this can shed more light on the implications of naturalistic metaethics for philosophical claims. They shouldn’t be regarded as conceptual a priori claims, but as hypotheses that need empirical confirmation. Naturalistic metaethics is not concerned with a priori conceptual necessities. It requires revising our concepts when they don’t fit into the best theories. In that sense empirical philosophers should be revisionists (see [Francén 2010](#), pp. 137 and 142 for a more detailed account of revisionism).

Before I go on, I want to offer one last thought about this. What might be the motivation for framing these positions as claims about conceptual necessity? [Roskies](#) writes:

I take it that internalist philosophers have intended to offer something stronger than contingent claims about human wiring (...) Only a view involving necessity or intrinsicality can distinguish moral beliefs and judgments from other types by their special content. (2008, p. 193)

But why do we need a priori conceptual necessities to distinguish between different kind of beliefs and judgments? We could start with very simple observations. Apparently people play a game of blaming and blessing: they use words like “good” and “bad” that are somehow different than other terms. The task of defining what morality is could be a descriptive anthropological enterprise. And I think this is in the spirit of naturalistic metaethics.

I have argued that it is enough for Prinz’s sentimentalism (and for internalism) to claim a strong Relation between moral judgments and emotions. But what kind of Relation is strong enough for it? A mere statistical connection is surely not enough. If the important part of the sentimentalist thesis is not the Modality of the whole claim, we have to analyze the terms “ne-

cessary” and “essential” in a non-modal way. One possibility, that harmonizes with naturalized metaethics, is to regard this connection as *functional*.⁵

In the next, and final, section I shall look at Prinz’s critique of non-cognitivism. I shall present a speculative alternative to his view that I hope, again, is in agreement with his proposal for a naturalized metaethics.

3 Defending non-cognitivism as an empirical theory

Here, I want to argue against Prinz’s attack on non-cognitivism. He thinks that there are good empirical reasons to reject non-cognitivism. His first argument is that cognitivism can predict the surface form of moral language better than non-cognitivism. First, I argue against this by pointing to non-cognitivist accounts of moral language that I think can predict this surface form. Second, I provide a speculative non-cognitivist theory of why moral language has the surface form we can observe. Again, I think my proposal is in agreement with Prinz’s naturalized metaethics. I do think, however, that it challenges him to explore the space of possible accounts. My proposal shows, I hope, that the empirical evidence cannot, at this point, decide this question.

The second question in Prinz’s decision tree is whether or not moral judgments are truth-apt. Can they be true or false? Theories that answer yes to this question are cognitivist, while theories that answer negatively are non-cognitivist. *Non-cognitivism* is a collective term that can refer to many different theories (Shafer-Landau 2003, p. 17). It consists of two theses (Roojen 2013, section 1.1): the first says that moral utterances do not express propositions; they’re not truth apt. This is a semantic thesis about moral language. The second thesis says that moral beliefs are not representational. They do not refer to anything in the world. This is a thesis about the mental state of the

moral agent. Here Prinz wants to defend cognitivism by providing empirically-informed reasons to reject non-cognitivism. He defines expressivism in the following way (we can think of Expressivism as one form of the first, semantic thesis of non-cognitivism):

Expressivism =_{Def} Moral assertions express mere feelings or non-assertoric attitudes, and do not purport to convey facts. (Prinz this collection, p. 7)

Prinz denies both of the two theses that make up non-cognitivism. He argues that the most obvious empirical prediction of non-cognitivism fails, as he thinks that if non-cognitivism was true we would expect our moral language to have a non-cognitive form (Prinz this collection, p. 16). But this is not the case. It seems that our moral language mostly has declarative form.

If this is correct, and if I don’t have reasons to disbelieve it, does it mean that non-cognitivism makes wrong predictions? I don’t think this is the case. Much of the work in non-cognitivism is dedicated to explaining this apparent tension. But I don’t think that this involves “elaborate logics”, as Prinz puts it (this collection, p. 16). Rather, most non-cognitivists provide theories about the nature of moral discourse that show that we should expect the surface grammar to be declarative. I don’t think that non-cognitivism has or needs to have these “obvious empirical predictions”.

The starting point is to look at the way language is used. It is not the literal meaning of ethical terms that are of interest but their *function* (Björnsson 2002, p. 328). Expressivism entails a pragmatist theory of moral language:

[T]he pragmatist attempts to describe the function that a word, phrase or concept plays in human life, and once he has satisfied his curiosity there, he does not think that there are any further questions to ask about utterances of that sort. (Smyth 2014, p. 608)

Arguably, such a pragmatist view is easier to naturalize because we have the social sciences,

⁵ For this proposal see Björnsson & Francén Olinder (2013) and Bedke (2009) and Schulte (2012). They detail the idea that we can think of this relation as *teleo-functionalistic*.

which offer large toolboxes for investigating human practices.

Although Prinz's definition of expressivism may be at the heart of non-cognitivism, in most cases this is not the whole story. According to expressivism, moral terms are not only used to express one's attitudes but also to provoke certain attitudes in the hearer. This idea goes back to the early emotivists. The "dynamic use" of language (Stevenson 1937, p. 21) involves the manipulation of others: "[E]thical terms are instruments used in the complicated interplay and readjustment of human interests" (Stevenson 1937, p. 20; emphasis in original).

Stevenson, and many others following him, analyze expressions like "x is good" as meaning "Hooray for x! Do hooray as well!" (Stevenson 1937, p. 25).⁶ It expresses the speaker's attitude and the wish or the prescription that the hearer should adopt this attitude as well.

At this point Prinz could reiterate his point and simply ask: "Why then do we say 'this is good' and not 'I like this, do so as well'?" Here I want to offer a speculative answer: because we don't like to be manipulated. If the function of moral language is, at least in part, to influence the attitudes and the behavior of others, I think we should expect it to take this form. This is because a declarative sentence has more *authority* than a mere expressive one. If I want someone to do something it is arguably more effective to disguise it in non-subjective form, to give it the appeal of a truth-aptness.⁷ I want to disguise it so that it will serve this persuasive purpose.

I don't want to say that these ideas are correct. But they're plausible theories that predict the surface form of moral language, and which are no worse than cognitivist theories. Expressivism focuses on what people do with language. It focuses on the speech act, not the literal meaning. Whether people express, declare, prescribe, describe, recommend, or evaluate is nothing we can easily read from the sur-

face form. But this is what Prinz seems to presuppose when he says the most obvious empirical prediction fails. We have to look at their behavior and the pragmatic context in which the discourse happens.

I argue that this fits even better with Prinz's project of a naturalized metaethics. When Prinz discusses the last step in the decision tree, he writes: "Naturalizing relativism will require the marriage of cultural anthropology and sociolinguistics" (this collection, p. 24). I think this marriage could be more helpful at an earlier stage in the decision tree—to help answer the question of whether or not moral terms aim at truth.

4 Conclusion

In this commentary on Prinz's highly interesting and substantial target paper I welcomed his methodological naturalism, but argued that his project is not as conciliatory between traditional analytical philosophy and naturalized philosophy as he seems to think. The reason is that on closer scrutiny we find opposing views on the methodology and purpose of philosophy. In the second part of my contribution I looked at an objection against Prinz's sentimentalism. I argued, first, that he misses the real core of this kind of objections. Then I used Jon Tresan's distinction between *de dicto*- and *de re*-internalism as a conceptual tool to propose and develop another answer that Prinz could use against this objection. In particular, I claimed that, given Prinz's metaethical naturalism, we should not look for conceptual necessity but for fruitful hypotheses which we can test in *a posteriori*. In the third and last part I argued against Prinz's critique of non-cognitivism. Prinz thinks that the most obvious empirical prediction of non-cognitivism fails. Here, I tried to demonstrate how non-cognitivism, given a pragmatical view of moral language, actually predicts the surface grammar of moral discourse as well as cognitivist alternatives. I proposed a speculative explanation for this interesting fact. This kind of explanation, I believe, fits even better with Prinz's project of a naturalized metaethics.

⁶ Stevenson (1937, p. 25) writes: "I do like this; do so as well!" But the first part looks suspiciously descriptive. Because this doesn't fit with Stevenson's account, I reformulated it in this way.

⁷ Mackie discusses this instrumental use when he discusses why people give their moral judgments the appeal of objectivity (1990, p. 42). But as we saw, Prinz thinks this premise is wrong.

References

- Bedke, M. S. (2009). Moral judgment purposivism: Saving internalism from amorality. *Philosophical Studies*, 144 (2), 189-209. [10.1007/s11098-008-9205-5](https://doi.org/10.1007/s11098-008-9205-5)
- Björklund, F., Björnsson, G., Eriksson, J., Francén Olinder, R. & Strandberg, C. (2012). Recent work on motivational internalism. *Analysis*, 72 (1), 124-137. [10.1093/analysis/anr118](https://doi.org/10.1093/analysis/anr118)
- Björnsson, G. & Francén Olinder, R. (2013). "Internalists beware" - We might all be amorality! *Australasian Journal of Philosophy*, 91 (1), 1-14. [10.1080/00048402.2012.665373](https://doi.org/10.1080/00048402.2012.665373)
- Björnsson, G. (2002). How emotivism survives immoralists, irrationality, and depression. *Southern Journal of Philosophy*, 40 (3), 327-344. [10.1111/j.2041-6962.2002.tb01905.x](https://doi.org/10.1111/j.2041-6962.2002.tb01905.x)
- Björnsson, G., Eriksson, J., Strandberg, C., Francén Olinder, R. & Björklund, F. (2014). Motivational internalism and folk intuitions. *Philosophical Psychology*, 1-20. [10.1080/09515089.2014.894431](https://doi.org/10.1080/09515089.2014.894431)
- Darwall, S. L. (1995). *The british moralists and the internal 'ought', 1640-1740*. Cambridge, UK: Cambridge University Press.
- DeLapp, K. M. (2011). Metaethics. *The Internet Encyclopedia of Philosophy*. <http://www.iep.utm.edu/metaethi/>
- Francén, R. (2010). Moral motivation pluralism. *Journal of Ethics*, 14 (2), 117-148. [10.1007/s10892-010-9074-y](https://doi.org/10.1007/s10892-010-9074-y)
- Kauppinen, A. (2008). Moral internalism and the brain. *Social Theory and Practice*, 34 (1), 1-24. [10.5840/soctheorpract20083411](https://doi.org/10.5840/soctheorpract20083411)
- Mackie, J. L. (1990). *Ethics : Inventing right and wrong*. London, UK: Penguin.
- Nichols, S. (2002). How psychopaths threaten moral rationalism: Is it irrational to be amoral. *The Monist*, 85 (2), 285-303.
- Prinz, J. (2006). The emotional basis of moral judgments. *Philosophical Explorations*, 9 (1), 29-43. [10.1080/13869790500492466](https://doi.org/10.1080/13869790500492466)
- (2008). Empirical philosophy and experimental philosophy. In J. Knobe & S. Nichols (Eds.) *Experimental philosophy* (pp. 189-208). Oxford, UK: Oxford University Press.
- (2015). Naturalizing metaethics. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Roskies, A. (2003). Are ethical judgments intrinsically motivational? Lessons from "Acquired Sociopathy". *Philosophical Psychology*, 16 (1), 51-66. [10.1080/0951508032000067743](https://doi.org/10.1080/0951508032000067743)
- (2008). Internalism and the evidence from pathology. In W. Sinnott-Armstrong (Ed.) *Moral psychology: The neuroscience of morality: emotion, brain disorders, and development* (pp. 191-206). Cambridge, MA: MIT Press.
- Schulte, P. (2012). Satan Und Der Masochist: Eine Nonkognitivistische Antwort auf den Amoralismus-Einwand. In A. Dunshirn, E. Nemeth & G. Unterthurner (Eds.) *Crossing Borders. Grenzen (über)Denken. Beiträge Zum 9. Internationalen Kongress der österreichischen Gesellschaft für Philosophie in Wien* (pp. 599-608). Wien, AUT: Österreichische Gesellschaft Für Philosophie.
- Shafer-Landau, R. (2003). *Moral realism: A defence*. Oxford, UK: Oxford University Press.
- Smyth, N. (2014). Resolute expressivism. *Ethical Theory and Moral Practice*, 17 (4), 607-618. [10.1007/s10677-014-9495-y](https://doi.org/10.1007/s10677-014-9495-y)
- Stevenson, C. L. (1937). The emotive meaning of ethical terms. *Mind, New Series*, 46 (181), 14-31.
- Strandberg, C. & Björklund, F. (2013). Is moral internalism supported by folk intuitions? *Philosophical Psychology*, 26 (3), 319-335. [10.1080/09515089.2012.667622](https://doi.org/10.1080/09515089.2012.667622)
- Tresan, J. (2009). Metaethical internalism: Another neglected distinction. *Journal of Ethics*, 13 (1), 51-72.
- van Roojen, M. (2013). Moral cognitivism vs. non-cognitivism. *The Stanford Encyclopedia of Philosophy, Winter 2013* E. N. Zalta (Ed.) <http://plato.stanford.edu/archives/win2013/entries/moral-cognitivism/>

Should Metaethical Naturalists Abandon *de dicto* Internalism and Cognitivism?

A Reply to Yann Wilhelm

Jesse Prinz

Yann Wilhelm pursues three issues in response to my target article. First, he tries to expose my naturalism as more radical than I let on. I concede the point, though I also offer ways in which my radicalism might be mitigated. Second, he exposes a limitation in my argument for internalism, and suggests that naturalists should defend form on internalism that is neutral about conceptual claims (*de re* internalism, rather than *de dicto*). I welcome the suggestion, but also consider how naturalists might defend *de dicto* internalism. Third, Wilhelm challenges my argument against non-cognitivism, by offering a novel explanation of the fact that moral judgments have an assertoric form. I response, I note avenues for cognitivist resistance to Wilhelm's explanation.

Keywords

Cognitivism | Conceptual truth | Internalism | Metaethics | Naturalism | Non-cognitivism

Author

Jesse Prinz

jesse@subcortex.com

City University of New York
New York, NY, U.S.A.

Commentator

Yann Wilhelm

ywilhelm@students.uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

In “Naturalizing Metaethics,” I try to establish that core questions in metaethics lend themselves to empirical investigation. I argue that we can potentially adjudicate long-standing debates by testing predictions made by competing metaethical theories. I also make some conjectures about how such empirical investigations will turn out. Based on a small selection of preliminary findings, I advance a case of a version of sentimentalism—the

view that emotions are essential to moral judgments. I also suggest that sentimentalism commits me to internalism—the view that moral judgments are essentially motivating—and I advance an empirical case for cognitivism—the claim that moral judgments are, like other assertions, capable of being true or false.

In his insightful commentary, Yann Wilhelm offers clarifications and challenges to my

arguments. First, he asks whether my naturalism is compatible with traditional approaches in philosophy. I imply that the two can co-exist in a complementary way, but Wilhelm suggests that my naturalism is more radical than it appears. I am forced to agree, and to clarify the co-existence claim. Wilhelm also challenges my case for internalism, distinguishing two different forms and suggesting I am only in a position to argue for one of them. I am open to that possibility, but I also sketch a strategy for defending both forms. Wilhelm concludes with a challenge to my defense of cognitivism. He provides non-cognitivists with an explanation for findings that I say they cannot explain. I offer a cognitivist response, but grant that this proposal demands empirical attention.

Wilhelm's commentary provides a valuable contribution to empirically oriented metaethics. He offers strategies for avoiding certain kinds of debates with opponents of naturalism, and he identifies empirical issues that can be used to settle debates between card-carrying naturalists. Wilhelm deepens my understanding of these issues and strengthens my optimism about the prospects of naturalistic metaethics.

2 Is naturalism a radical position?

Before moving on to the first order debates that Wilhelm so helpfully pursues, I want to concede an important point that he makes in the opening of his commentary. Wilhelm rightly observes that I overstate the extent to which a thoroughgoing naturalism can preserve traditional approaches to philosophy. Though ostensibly a plea for conciliation, I am, in fact, skeptical about the notion *a prioricity*. Rather, I claim that armchair methods are observational (intuitions are defeasible inner observations informed by prior experience, and open to empirical correction). As Wilhelm makes clear, traditionalists who view conceptual analysis as an *a priori* endeavor will not share my enthusiasm for naturalism.

In another respect, however, my position is conservative. I don't think traditional philosophers must stop working as they currently do. Armchair methods remain the primary source of

philosophical theories and distinctions. They also are the primary source for philosophical thought experiments that can be used to test between theories. Thus, my invitation to interpret armchair methods as observation is intended as a vindication of traditional philosophy, though not a vindication of how some traditional philosophers understand their own endeavors.

Proof of this qualified vindication comes from the fact that empirically oriented philosophers regularly draw on traditional work in devising their studies. For example, experimental philosophers have used trolley problems, twin earth cases, and the thought experiments used to back contextualism in epistemology. In my target paper, I relied on theories that have been identified and articulated within traditional philosophy. Testing between theories requires observation, I believe, but it would be a great loss if every philosopher ran a laboratory. Instead, I envision a future for philosophy in which many researchers do no experimental work, others are primarily experimentalists, and still others do a combination of the two. If we begin to make empirical methods a standard part of philosophical training, then philosophers will be able to read psychological research more responsibly and conduct experiments when they see fit. But it doesn't mean that they will also suddenly stop thinking and blindly collect data. As in the sciences, theoretical work is required in philosophy. We can resist the idea of *a priori* truth without throwing away the armchair.

3 Must naturalist be content with *de re* internalism?

These methodological points bear on Wilhelm's first challenge to my metaethical conclusions. In the target paper I argue for a form of internalism (roughly, the view that moral judgments are essentially motivating). Wilhelm points out that my evidence for this claim will not satisfy many externalists. I primarily rely on evidence that moral judgments always co-occur with emotional states, but for externalists will be impressed; they will say that such findings cannot address questions about whether it is necessar-

ily the case that moral judgments are motivating, even if they always happen to be motivating.

Wilhelm helpfully replies to this objection on my behalf, using Jon Tresan's distinction between *de dicto* and *de re* internalism. The former is a thesis about the concept of moral judgment (viz., it is a conceptual truth that when that concept applies, motivation applies as well). The latter is a claim about moral judgments themselves (viz., moral judgments do in fact carry motivation force). Wilhelm concurs that my evidence can contribute to a defense of the *de re* claim. He suggests that I abandon the case for *de dicto* internalism, since naturalists should not concern themselves with conceptual claims.

I welcome Wilhelm's suggestion, and I am inclined to endorse it. Let me mention, however, a strategy available to the naturalist whose heart is set on defending the *de dicto* claim. Returning to Wilhelm's discussion of methodology, let's imagine that naturalists wage a successful campaign against the *a priori*. Properly pursued, such a campaign might also undermine metaphysical necessity. Metaphysical necessities, unlike nomological necessities, are alleged to be true in virtue of conceptual entailments rather than laws of nature or natural facts. The critique of *a prioricity* threatens metaphysical necessity because it advances the view that truths about concepts are open to empirical revision. Let's suppose that concepts are mental representations garnered through experience with the function of classifying things in the world. So construed, concepts are susceptible to improvement through empirical inquiry. Initial concepts are rough and ready pointers that we use to carve up the observational world, and revised concepts are carvings that remain after observation. Now let us define a "robust conceptual truth" as the conceptual entailments that survive after a concept has been subjected to empirical fine-tuning. Such truths would more or less coincide with how the world is, together with certain pragmatic assumptions that go into theory construction. Thus, they would coincide with truths that emerge from our study of the things themselves (which are also constrained

by pragmatic assumptions). On this picture, *de dicto* collapses into *de re*. A defense of *de re* internalism would indicate that our concept of moral judgment will converge on internalism as well. Rather than bypassing *de re* internalism, we can try to defend it by naturalizing conceptual truth.

Wilhelm might reply that this defense of *de dicto* internalism would not persuade non-naturalists. The defense is based on the assumption that the naturalist critique of *a prioricity* goes through, but that is just what non-naturalists are inclined to deny. Thus, it might appear that the debate over the *de dicto* position is hostage to unresolvable disputes about the nature of philosophy.

Here I'd balk at the claim that such disputes are unresolvable. Those who believe in *a prioricity* may dislike naturalism, but they certainly believe that their views require evidential support. Naturalists offer an account of what concepts are (mental representations) and an explanation of conceptual intuitions (introspection of mental representations). Non-naturalists are obliged to provide an alternative account of both, and the two accounts can then be compared by agreed upon standards. I venture that the naturalist account will find a resounding victory in such a head-to-head match. It is more parsimonious view, since both sides must grant the existence of mental representations, and I suspect it can fully account for our conceptual intuitions.

These are, of course, big debates, which I cannot settle here. My point is simply that we can imagine a two-stage process that begins with broad issues about naturalism, and then moves on to first-order views. On my prognosis, we won't end up abandoning the notion of conceptual truth, but rather revising it. If so, *de dicto* naturalism might turn out true. Wilhelm may be right, however, that until we come to greater consensus on the nature of philosophy, naturalists might be on firmer ground if they try to bypass conceptual questions. He is also right that, from a naturalist perspective *de re* internalism may be the more interesting thesis. Conceptual claims lose their distinctive interest if concepts are revisable and, ultimately, coincident with empirical theories.

4 Can non-cognitivists explain the assertoric form of moral judgments?

Let me turn, finally, to Wilhelm's constructive effort to defend non-cognitivism. Non-cognitivists claim that moral judgments are not like ordinary assertions; they cannot be assessed as true or false, but rather merely express the speakers attitudes and commendations. If so, I ask, why do we express moral judgments as assertions? This is a familiar challenge. In my discussion, I merely point out that can be backed up by empirical data. Wilhelm has a two-part reply. First, he observes that, for non-cognitivists, the primary function of moral discourse is to persuade. C. L. Stevenson, for example, says that "x is bad" does not just mean "boo to x!"; it also means and "say boo to x as well!". Second, Wilhelm makes the original and plausible suggestion that this persuasive function is most effective when it covert. People, he notes, don't like to be manipulated. If I explicitly exhort you to say "boo!" you may resist, because no one likes being told what to do. But if I present my attitude in the form of an assertion, you might causally take it on board, as you would if I were presenting an ordinary statement of fact.

I think Wilhelm's proposal deserves serious exploration. Cognitivists can respond in two ways. First, they can try to show that moral discourse often occurs in contexts that don't aim at persuasion. This might seem implausible. After all, why should we bother engaging in moral discourse if we don't intend to persuade anyone? On closer analysis, however, it does seem that much of our moral discourse involves preaching to the choir. In political debates, for example, left wing pundits and right wing pundits engage in a lot of moral discourse, but they never seem to persuade each other. This raises the intriguing possibility that moral judgments are not primarily in the business of persuasion. An alternative possibility is that we make moral judgments to assert our identity, or express solidarity with like-minded individuals. Empirical tests might be designed to compare the persuasion model and the self-expression model.

Cognitivists might also try to resist Wilhelm's conjecture that people do not like to be manipulated by consulting research on explicit persuasion. In defense of Wilhelm's conjecture, there is a literature suggesting that people sometimes resist explicit persuasion (e.g., [Petty & Cacioppo 1979](#)). On the other hand, resistance does not occur in all contexts. Indeed, in a consumer product context, [Reinhard et al. \(2006\)](#) found that, when a person is regarded as likeable (or attractive!), they become even more persuasive when they make their intent to persuade explicit. Similarly, in studies of college drinking behavior, [Neighbors et al. \(2008\)](#) found that injunctive norms (which explicitly reference attitudes) are effective when and only when they are expressed by members of the students' social groups. Further work could test the effects of explicit injunctions in the moral domain.

I should underscore that I think more testing is required to settle these debates. Wilhelm's explanation for surface discourse remains viable, and we can make progress on these issues by devising new ways to test it. These are manifestly empirical issues. While I wager with the cognitivists, I grant that the case is far from closed.

5 Conclusion

I am indebted to Yann Wilhelm for his generous and probing commentary. It brings welcome clarification and new challenges to the project I set out "Naturalizing Metaethics." I also welcome the spirit of Wilhelm's discussion, which moves beyond ideological debates about metaphilosophy, and offers promising strategies for answering core metaethical questions.

Wilhelm successfully establishes that my preferred form of naturalism is less compatible with traditional philosophy than I let on, but I also pointed out that work by traditionally minded philosophy remains an invaluable font of philosophical theories. Wilhelm then offers a helpful suggestion that naturalists might more easily defend internalism if they bypass conceptual versions of that view. In response, I suggested that the radical implications of naturalism may actually offer a way to defend the concep-

tual version of internalism, by advancing a naturalized account of conceptual truth. Finally, Wilhelm offered a new psychological cum functional account of moral discourse, which inoculates non-cognitivists against grammatical objections. While I hold out hope for cognitivism, Wilhelm has identified a genuine empirical challenge to the cognitivist. This challenge beautifully demonstrates the value of empirical testing in metaethics, and it also reminds us that there is much work to be done.

References

- Neighbors, C., O'Connor, R. M., Lewis, M. A., Chawla, N., Lee, C. M. & Fossos, N. (2008). The relative impact of injunctive norms on college student drinking: The role of reference group. *Psychology of Addictive Behaviors: Journal of the Society of Psychologists in Addictive Behaviors*, 22 (4), 576-581. [10.1037/a0013043](https://doi.org/10.1037/a0013043)
- Petty, R. E. & Cacioppo, J. T. (1979). Effects of forewarning of persuasive intent and involvement on cognitive responses and persuasion. *Personality and Social Psychology Bulletin*, 5 (2), 173-176. [10.1177/ 014616727900500209](https://doi.org/10.1177/014616727900500209)
- Reinhard, M.-A., Messner, M. & Sporer, S. (2006). Explicit persuasive intent and its impact on success at persuasion – The determining roles of attractiveness and likeableness. *Journal of Consumer Psychology*, 16, 249-259. [10.1207/s15327663jcp1603_7](https://doi.org/10.1207/s15327663jcp1603_7)

The Representational Structure of Feelings

Joëlle Proust

The word “feeling” denotes a reactive, subjective experience with a distinctive embodied phenomenal quality. Several types of feelings are usually distinguished, such as bodily, agentive, affective, and metacognitive feelings. The hypothesis developed in this article is that all feelings are represented in a specialized, non-conceptual “expressive” mode, whose function is evaluative and action-guiding. Feelings, it is claimed, are conceptually impenetrable. Against a two-factor theory of feelings, it is argued, in the cases of affective and metacognitive feelings, that background beliefs can circumvent feelings in gaining the control of action, but cannot fully suppress them or their motivational potential.

Keywords

Affective feelings | Affordance | Agentive feelings | Appraisal | Arousal | Bodily feelings | Comparator | Control | Cues | Evaluative | Expressive | Familiarity | Fluency | Formal object | Illusory feeling | Incidental feelings | Integral feelings | Intensity | Metacognitive or noetic feelings | Monitoring | Nonconceptual content | Predictive | Reactive | Resonance | Retrospective | Somatic marker | Transparency | Two-factor account | Valence

Author

Joëlle Proust

joelle.proust@ehess.fr

Ecole Normale Supérieure
Paris, France

Commentator

Iuliia Pliushch

pliushi@students.uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

“Feeling” denotes a reactive, subjective experience with a distinctive embodied phenomenal quality and a formal object, which may or may not coincide with embodied experience. Feelings typically express affect and valence in sensation. “Reactive” means that feelings are closely associated with an appraisal of a present property or event. The term “reactive” is crucial. The term “feeling” is sometimes used to refer to a non-reactive, perceptual experience. For example, when one perceives an object through touch, it is common to say that “one feels one’s key in one’s pocket”. But “feeling”, in this context, does not refer to a reactive phenomenon.

It rather refers to the feedback of one’s own key-touching activity. This type of perceptual feeling is expected to result from one’s action and, hence, does not belong to the domain of reactive feelings. What is called the “formal object” (see [Kenny 1963](#)) of a feeling is the property in the triggering event that elicits the reactive feeling. For example, the formal object of fear is some threatening property detected in the perceptual field.

Feelings can be pleasant or aversive, strong or weak, short-lived or long-lasting, or have an arousing or depressing character. They motivate distinctive dispositions to act, whose

Glossary

Feeling	“Feeling” denotes a reactive, subjective experience with a distinctive embodied phenomenal quality and a formal object, which may or may not coincide with the embodied experience. Feelings typically express affect and valence in sensation.
Reactive	“Reactive” means that feelings are closely associated with an appraisal of a present property or event.
Formal object	“Formal object” of a feeling is the property in the triggering event that elicits the reactive feeling.
Metacognitive feelings	Metacognitive feeling are experienced while conducting a cognitive task: the agent may find the task easy or difficult, anticipate her ability or inability to conduct it. Once the task is completed, the agent may have the feeling of being right, or have a feeling of uncertainty about the outcome of her endeavour.
Affordance	Affordances are positive or negative opportunities, expressed in feelings: an affordance-sensing swiftly and non-reflectively motivates the agent to act in a particular way.
FS Affordance	FS Affordance _a [Place _a =here], [Time _a =Now/soon], [Valence _a =+], [Intensity _a =.8 (comparatively specified on a scale 0 to 1)], [motivation to act of degreed according to action programa].
Transparency	A mental state is transparent if, when it is activated, its intentional content is accessible to the subject who entertains it.
Incidental and integral feeling	Metacognitive feelings are called “incidental” when they are not based on valid cues about the cognitive task at hand, and hence, have no predictive value. They are called “integral” when they actually carry information about cognitive outcome.

urgency is entailed both by the feeling experience and the context in which it is experienced: feeling an intense pain disposes the person to promptly locate and remove the cause of the pain; except, for example, when it is self-inflicted, or when it is part of a ritual.

Most theorists of feelings agree that they are associated with—or, for those who identify emotions with conscious experiences¹ consist of—specialized, internally generated bodily sensations, such as an increase in heart rate, contractions or relaxations of the facial muscles, visceral impressions, tremors or tears, impulses to run away, etc. As will be seen below, some feelings, however, do not express emotions., i.e.,

¹ From the viewpoint of the somatic feeling theory of emotions, emotions can be explained as a somatic change caused by the perception, real or simulated, of a particular object. See [James \(1884, p. 190\)](#), and [Damasio \(1994, 2003\)](#). Other theorists of emotion, however, consider that the conscious experience of having an emotion includes propositional attitudes, and not only feelings. See sections 4 and 5 below. Moods are long-term affective states, and will not concern us here.

they are not affective. A feeling tends to be more explicitly felt as bodily when it has a body-related function; that is, the phenomenology makes the need to be served salient (feeling tired, feeling a pain in the joints) in order to motivate action. In affective feelings, in contrast, the bodily phenomenology tends to recede to the fringe of consciousness (feeling in love with *A*, feeling angry with *B*).² From this observation, it is easy to infer that types of feelings differ in their respective meanings: they in some sense *express what they are about*. In affective feelings, an experience of “feeling toward” is supposedly present: the emotion is felt as being about an object, a person, or a situation—the objects, rather than bodily sensations, are the focus of one’s emotional attention. Affective feelings also include mixed cases where one seems to both experience a strong bodily feeling at the same time as the intentional content that

² On this concept, see [Mangan \(1993, 2000\)](#) and [Reber et al. \(2002\)](#).

this feeling seems to refer to, as when Marcel Proust's narrator reports experiencing an acute pain in the chest when thinking about his beloved deceased friend, Madame de Guermantes.³ It is unclear whether metacognitive (also called noetic, or epistemic) feelings are affective or non-affective (see section 7 below). They are experienced while conducting a cognitive task: the agent may find the task easy or difficult, and may anticipate her ability or inability to conduct it. Once the task is completed, the agent may have the feeling of being right, or may have a feeling of uncertainty about the outcome of her endeavour. Take the case of a person who feels unable, presently, to remember what she had for dinner last night. Her feeling of not remembering is correlated with activity in a facial muscle, the corrugator supercilii (Stepper & Strack 1993). Her feeling, however, is not about her disposition to contract or relax this or that muscle, of which she is certainly unaware. It is, rather, about her present disposition to remember what she had for dinner. Epistemic feelings seem to be “feeling-toward” experiences, and have cognitive dispositions or contents as their object.

Descriptive phenomenology, however, does not offer in itself an account of the intentional structure of feelings. We need to understand how feelings in general gain their real or supposed aboutness, and how they relate to action-guidance as a function of context; i.e., we need to provide a functional analysis of feelings. Section 2 will begin to provide such an analysis, and will address a preliminary issue—namely, Do the phenomena that are usually called “feelings” share a property that makes them a natural kind? In section 3, the specific informational structure of feelings will be seen to account for their generic characteristics. Section 4 will clarify the account by way of addressing various objections. Section 5 will attempt to show that the proposed account fares better with experimental evidence than a cognitivist account of affective and metacognitive feelings. Section 6 will examine whether or not metacognitive feelings have an affective valence.

2 Are feelings a natural kind?

Paul Griffiths has claimed that *emotions* do not constitute a natural kind, in the sense that they do not form “a category about which we can make inductive scientific discoveries” (2004, pp. 901–911). One can agree with latter claim, however, without concluding that *feelings* do not constitute a natural kind. First, feelings are not only affective ingredients in emotional awareness. Some feelings, such as feeling cold or sick, or feeling that one is acting, have nothing to do with affective episodes. Second, there are evolutionary reasons to distinguish, within emotions, two classes of subjective appraisals. Emotion theorists usually contrast *feelings* expressed in primary emotions—fear, anger, happiness, sadness, surprise, and disgust—with *various appraisals cum conative dispositions* associated with higher cognitive emotions, such as envy, guilt, pride, shame, loyalty, vengefulness, and regret. The first are phylogenetically and ontogenetically prior to cognitions. They belong to the ancient limbic system, which is present in some form in most animals. A quick route from the retinal image to the amygdala through the thalamus allows affective information to control behavior (see LeDoux 1996). Primary feelings are thus triggered independently of concept possession and motivate specific responses. Secondary affective experiences, in contrast, might have evolved on the basis of social constraints in relation to cooperative action among humans. Indeed (with the possible exception of pride and shame) they are not present in nonhuman *primates*.⁴ They activate newer brain structures; they require concept possession, depend on background beliefs, and do not generate characteristic behaviors. Finally, primary feelings are clearly embodied, while secondary emotions seem to have no proprietary somatic markers. An interesting idea, suggested by Jesse Prinz (2004, p. 95), is that the facial or somatic correlate of secondary emotions, *when they have one*, involves a blend of the somatic markers for primary feelings.

³ See the analysis of this example in Goldie (2002), p. 56.

⁴ On this contrast, see Frank (1988), Griffiths (1997), and Prinz (2004, pp. 82–83). On whether they qualify as emotions, see Ekman (1992).

In summary: emotions differ, among other things, because of the unequal role that feelings have in the two classes of emotions just discussed. The wider scope of feelings, when understood as “reactive, subjective experiences with a distinctive embodied phenomenal quality”, seems to be more unified than emotions, and making feelings seem like plausible candidates for a natural kind.

We need, however, to turn this tentative definition into a general functional characterization that presumably holds for all feelings (beyond affective ones) and only for them. Here is a proposal: feelings constitute the sensitive part of predictive and retrospective processes of non-conceptual evaluation of one’s own and others’ well-being and actions. Being essentially evaluative, feelings are always the output of a comparator: in other terms, they are crucial *monitoring* ingredients in self-regulated adaptive control systems. In such systems, the specific function of a feeling consists in detecting how much a current *observed* value of a parameter *deviates from its expected value*, on one or several dimensions relevant to survival (see Carver & Scheier 2001). Their formal object, when they have one,⁵ (such as being afraid of the bear in front of me) cannot be analyzed independently of the monitoring function they serve within a specialized control loop.⁶ Relevance to well-being, however, extends to bodily condition, goal achievement, and availability of preferred goods of all kinds (food, partner, social status). The relevant dimensions of variation that feelings track may accordingly be of a sensory, proprioceptive kind (feeling thirsty, cold, etc.), social-affective (feeling angry), or agentive (goal-related). Goal achievement, however, involves either epistemic or instrumental success, respectively generating epistemic feelings (feeling interested, bored, epistemically uncertain) and agentive feelings (feeling of happiness, of agentive confidence, of ownership of one’s action, etc.). Feelings, in summary, are the outcomes of comparators in a con-

trol loop; they carry non-conceptual information about how much one’s present condition deviates from the expected condition. From a functional viewpoint, they form a natural kind insofar as their function is to indicate a comparative outcome through a dedicated embodied experience.

Note, however, that there are comparators that trigger no feelings at all: these non-sensitive comparators may either work outside consciousness (for example, error signals driving immediate correction⁷, not to mention comparators that work at the cell level), or they can take concepts as their input, rather than reacting to percepts or situations (for example comparators of currency or of educational value).

As far as feelings are concerned, they are directly related to a presently-perceived context (or an imagined or remembered context, but in a “present-like”, indexical mode): one can feel too hot, too cold, or too tired (or feel “OK”, which usually means a tolerable deviation from the expected value). One can feel the fright one has had, even after the frightening event has ended. The outcome of a feeling-based appraisal, from a functional viewpoint, has to consist in some disposition to act that is adaptive, relative to the input to which the feeling is a reaction. Granting that feelings, as sensitive comparators in a control system, form a natural kind, there should be common properties cutting across the various types listed above. In fact we find three types of functional relations between feelings of a given kind and the associated disposition to act. First, feelings, according to their embodied valence, typically determine actions of approach or of avoidance. Some dictate caution, others boldness. Some encourage self-restraint, others self-assertion. Fear promotes a flight tendency, hunger a tendency to approach food. Second, they have a specific orientation in time: some feelings have a *predictive* function, and thus induce a behavior that is based on contingencies to be *further* displayed in the present context. For example, fear, when directed at a possible danger, increases the readiness to flee in case the danger concretizes.

⁵ As observed by Goldie (2009), some feelings, for example, [feeling anxious] or [feeling depressed], seem to lack a formal object, which is typically the case with moods. As indicated above, moods will not be discussed in this article.

⁶ Bechara et al. (2000) make it clear that the somatic marker theory applies to action, whether it engages affects or not.

⁷ see Logan & Crump 2010 and Nieuwenhuis et al. 2001

Others have a *retrospective* function, and induce corrections to the commands one has previously used, or to one's previous preferences. For example, feeling nauseous *after* food ingestion induces food avoidance, i.e., a change in the agent's preferences. In contrast, feeling disgust at the sight of some food may prevent the agent from approaching it. A subset of feelings, such as feeling happy, have both temporal orientations. Third, according to their embodied dynamics and intensity (which is called their "level of arousal"), feelings can provoke an elevation in the energy available to the system: they provoke excitement, agitation, power in the coming response; or, on the contrary, they may have a soothing effect and diminish the tendency to act.

One major functional property of feelings, from the viewpoint of information extraction and use, is that they can very rapidly extract and synthesize multiple cues from perception. This rapidity is a consequence of the automatic and encapsulated character of the control mechanism whose output they express. Feelings are automatically triggered by a specific type of input (which is the definition of informational encapsulation).⁸ Automaticity is associated with feelings being inescapable, at least for those feelings that have been allowed to develop within a culture, granting normal development.⁹ The mechanism that generates somatic, noetic, or affective feelings from inputs (perceptual, imaginative, or memorial) does not require one to have specific beliefs or intentions.¹⁰ Informational encapsulation explains why transitive feelings persist when the agent finds out that the situation is different from what she thought to be the case. Just as an optical illusion such as the Müller-Lyer effect does not immediately dissipate when it turns out that the segments are equal, a feeling of anger does not disappear as soon as the agent realizes that its formal object is not exemplified.

⁸ Automaticity in appraisal is central to Ekman's analysis of primary emotions (1992). See also Griffiths (1997), Prinz (2004), and Zajonc (1980). On informational encapsulation, see Fodor (1983).

⁹ For example, fearlessness in the presence of danger may result from a disturbed childhood.

¹⁰ Some affective feelings, however, can be intentionally controlled in the long run, through cultural learning. See Murata et al. (2013).

Automaticity and informational encapsulation seem also to characterize agentive feelings (see Pacherie 2008). Feelings generated in the course of a physical action come in two varieties: generalized or specialized. Some, such as feelings of agency, of initiation of action, of ownership and of motor control, are indicators monitoring action in progress: they concern "who" is performing the action, and "how" the action is being conducted (see Proust 2000). Others concern the evaluation of an action in one's own repertoire: a professional carpenter or an experienced musician, for example, have feelings telling them if an action sequence (whether their own or another agent's) in this repertoire sounds or looks right, even before they identify why they have this feeling. These feelings are also the outputs of a comparison between motor anticipations and observed properties of the action (a "forward model of action" supposedly stores the expected values of crucial parameters; Wolpert et al. 2001). They can predict the likelihood of completing an action (when the question arises, in difficult or non-routine cases), or evaluate—on-line or in retrospect—how swiftly, effortlessly, or unhesitatingly an action was performed. Agentive feelings thus have an essential role in regulating the fundamental properties of physical actions, such as the quality of the outcome,¹¹ and the ownership of the action.¹²

Noetic feelings, finally, are functionally similar to somatic, affective, and agentive feelings—although their evolutionary pattern seems to be different from the other three kinds. While most organisms have proprioceptive, affective, and motor control, and hence, presumably, somatic, affective, and agentive feelings, few are able to control their cognitive decisions through metacognitive feelings (see Beran et al. 2012 and Proust 2013). The latter are generated when trying to perceive, to remember, or to plan a cognitive task (in particular, when trying to plan how long to study material in order to

¹¹ Non-conscious error signals can also guide corrective steps, without the agent noticing them.

¹² Pat Haggard et al. (2002) have demonstrated the crucial role of the temporal binding between felt initiation of action and output in the sense of being the agent of an action. See, among other articles, Haggard et al. (2002).

master it).¹³ They are also relied upon when trying to reason or to solve a problem; when conversing, feelings of effort, and of informativeness, are monitored by speakers and hearers in order to maintain a common level of relevance. Like other feelings, they have two distinctive temporal orientations. Some have a predictive function. A feeling of knowing (FOK) may arise when trying to remember an item—for example a proper name—that one has not yet retrieved: having a strong FOK reliably predicts that one will finally retrieve the searched content (Koriat & Levy-Sadot 2001). A feeling of having a name on the tip of one's tongue (TOT) both signals the fact that a word is not presently available, and, according to its onset, valence, and intensity, whether it is worth or not worth pursuing one's effort to retrieve it (see Brown 1991 and Schwartz et al. 2000). Feelings of fluency are the sense of ease of processing one may feel or fail to feel when attempting to perceptually discriminate objects with a given property, or to retrieve items from episodic or semantic memory. A feeling of familiarity is particularly salient, in human adults, when no further fact about the target can be retrieved. It offers useful information about the epistemic status of the target: that it is not new, but nevertheless not fully recognized. A feeling of familiarity, then, motivates, among others, an attempt to recognize what or who a target is. Other metacognitive feelings have a retrospective function. When a name is retrieved, a feeling of rightness (FOR) motivates the agent to consider her response the expected one.¹⁴ Various feelings of uncertainty, based on fluency, coherence, plausibility, informativeness, or relevance, also have retrospective functions: their valence and intensity tell the agent whether she should accept or reject a cognitive outcome. These parameters are expressed through specialized somatic markers, such as increased activity in the facial muscle involved in smiling, the zygomaticus major—for positive valence—or the corrugator supercilli (involved in frowning)—for negative valence (Winkielman & Cacioppo 2001).

Taken together, these considerations are compatible with the view that somatic, agentive, metacognitive, and “primary” affective feelings, even if they differ in their formal objects, form a natural kind. Our attempt above at a functional characterization focused on the general relations of feelings to inputs, outputs, and mediating evaluative mechanisms. From this characterization, it emerges that feelings are gradients in comparators that are felt subjectively, rather than being propositional states describable in analytic, objective terms. These observations, however, suggest that, in order to express a specialized and fine-tuned reactivity to one or several formal objects, and to motivate adapted behaviors, in order to be remembered and conveyed to others feelings must have their own representational format. We now turn to the following question: What is the structure of the information that is extracted and expressed in a feeling?

3 What kind of information do feelings express?

The above question is important for clarifying the relation of feelings both to their formal object, when they have one, and to the action that they motivate. In the case of metacognitive feelings (M-feelings), the difficulty is particularly pregnant: it stems from the fact that, if we grant that M-feelings do not require concept possession to be felt, then it is unclear how their formal object should be construed: What are they about? Let us take a feeling of uncertainty, felt while trying to remember a proper name. Is this feeling about a memory *state*, or about a *disposition* to retrieve a proper name? If a feeling is about a memorial state or a disposition, its intentional content needs to include concepts of memory, of correctness, and of uncertainty. Empirical evidence, however, demonstrates that animals with no mindreading ability, and hence that are deprived of concepts of perception or of memory, are able to monitor their perception and memory as reliably as humans do.¹⁵ Furthermore, human children, from

¹³ This prediction involves judgments of learning (JOL). See Koriat & Ackerman (2010).

¹⁴ On FORs, see Thompson et al. (2011).

¹⁵ Rhesus monkeys have been found to opt out of more or less challenging perceptual or memory trials as a result of trial difficulty. For a

early on, are sensitive to the contrast between familiar and unfamiliar faces and environments. This supports our claim above: one can feel cold or anxious or uncertain without having the corresponding concepts of those feelings. A propositional format does not seem to apply to feelings in general.¹⁶

How do feelings fulfill their particular embodied, subjective way of representing—a mode we will call the “expressive mode”? The broadly functional characterization given above provides useful clues. Expressive representations comprise exclusively non-conceptual, perceptual, and evaluative (gradient- and valence-based) elements, which taken together express a subjective relation to the environment (internal or external) and a given tendency to act. It should be emphasized, however, that adult humans can obviously entertain *simultaneously* expressive and conceptual representations. The present hypothesis, in conformity with the literature on dual-processing, is that the expressive system processes information and influences decisions on the basis of its own narrow range of associations and norms; while the conceptual system takes advantage of background beliefs and inferential reasoning to make decisions in light of a broader set of norms. Let us take the case of an agent feeling joy after having won the lottery. A human adult normally has [lottery] in her conceptual repertoire, along with some of the inferences that can be made on its basis. However, the agent’s reactivity to the winning event falls under the expressive mode of representation, because this is the mode in which evaluation of the opportunities is conducted. This feeling representation presumably enlightens and orients the concept-based reasoning that can be conducted concerning the same event, such as wondering how to spend the money, or whether quitting her job is a good idea. We propose to call “affordance-sensing” the information that a feeling expresses. Affordances are positive or negative opportunities, expressed in feelings: an affordance-sensing swiftly and non-reflectively

motivates the agent to act in a particular way. Departing somewhat from Gibson’s use of this term within his ecological theory of perception, “affordance” is used here to refer to a non-conceptual and entirely subjective appraisal of the environment by the agent: an affordance is a perceived utility, which can be positive (something to approach and grasp) or negative (something to avoid and from which to flee).¹⁷

The corresponding representation has an indexical structure, because it has an essential relation to an occurrent represented property. Indexicality, however, has to be understood here in a non-referential sense. What is indexed is an *occurrent* (relational) affordance, rather than an individual event or object. Here is our proposal for what a given feeling structure (FS) looks like:

- FS Affordance_a [Place_a=here], [Time_a=now/soon], [Valence_a=+], [Intensity_a=s (comparatively specified on a scale 0 to 1)], [motivation to act of degree_a according to action program_a].

The subscript “_a” is meant to indicate that all the elements that have this subscript are representational cues, i.e., ingredients, in present affordance-sensing *a*. Note that the strength (or degree) of the motivation to act does not depend only on the fitness significance, i.e., on the valence and intensity of the affordance. Other factors, such as the physical condition of the agent and her prior arousal level (her mood) also modulate her motivational level (Schwarz & Clore 2007). The specification of the location of the affordance may vary, depending on the way the feeling was generated, but indexicality and reactivity suggest that the relevant affordance is often sensed to occur where the feeling is experienced. As will be seen later, however, M-feelings do not involve a specification of place.

The feeling structure proposed above includes somatic markers, even if they are not

summary of the results and a methodological discussion of their significance, see Beran et al. (2012), Chapter 1, and Proust (2013), Chapter 5.

¹⁶ For a defense of emotional representations as nonconceptual and action oriented, see Griffiths & Scarantino (2009).

¹⁷ See Proust (2009, 2013). Prinz (2004) briefly discusses this idea in connection with the intentional content of emotions (p. 228). See also Griffiths & Scarantino (2009): in emotion, “the environment is represented in terms of what it affords to the emoter in the way of skillful engagement with it.”

made explicit: these markers are the substrates for the information of valence and intensity. This information is carried by neural activations and associated bodily changes, such as a sudden sensation of pleasant muscle relaxation, or of unpleasant muscle contraction, or of visceral contractions associated with fear. Intensity of affordance, i.e., the arousal produced by a feeling, is also felt through the comparative amount of bodily reactivity to the affordance. These somatic markers, as emphasized above, are themselves part of a monitoring system designed to predict and assess one's relations to the environment along the relevant dimensions listed above (agency, individual and social well-being, preferences, and metacognition).

Let us consider further how to read the feeling structure given above. It is meant to reflect not only what is presently felt, but also what is stored in memory when a feeling is experienced, what can be imagined, and what can be conveyed to others in expressive behavior. The central idea is that feelings sensitively express *a subjective, embodied relation to an opportunity* in an input from the environment (understood in a broad sense as including external and bodily properties relevant to well-being). This primitive intentional relation is best captured by the term *affordance-sensing*. Feelings express this affordance as their focus (or formal object), along with its graded valence—ranging from very unpleasant to very pleasant—and with its intensity gradient, which ranges from small to large.¹⁸

As often emphasized, reactivity to an affordance occurs very rapidly in a processing sequence—even before the perceptual processing has been completed—and well before a concept-based judgment can be made (see Dolan 2002, p. 1191; Griffiths 1997, pp. 77; LeDoux 1996, pp. 174; Prinz 2004, pp. 60, and Zajonc 1980, pp. 153). This suggests that an alternative,

evaluative informational system screens the input with its own independent memorial structures.¹⁹

An affordance does not need to have an objective counterpart to be sensed, i.e., for a feeling to arise: it is enough that the agent anticipates it (even wrongly), imagines it, or remembers it, for the corresponding feeling to be expressed. A feeling, thus, does not presuppose a conceptual appraisal of the context, but rather it indexes in an embodied way a direct evaluative registration. Given that an affordance does not aim at characterizing the world, one cannot say, when the expressed affordance has no objective counterpart, that a feeling “misrepresents” the world as having a given affordance, or reciprocally that an existing affordance was “missed” by the agent when the latter failed to detect it. For misrepresentation to occur, a system must be equipped to attribute properties to individual objects, that is, it must be able to apply concepts. The expressive system, however, does not refer to objects as independent entities. Hence, affordance is not literally what a feeling is about, because aboutness presupposes that what is represented is independent from the representational system. Being relational, affordances cannot be grasped independently of the experience of a sensitive agent. When saying that a feeling “expresses” an affordance, we mean that it “resonates” to it (or that it monitors it). Resonance is a neural-somatic reactivity: it carries indexical and evaluative information, but it does not refer to the world or attempt to describe it.

It is possible, however, to objectively characterize what a feeling functionally refers to, and to pinpoint cases of misrepresentation, by re-describing the feeling structure above in non-subjective, non-evaluative propositional terms. Taking advantage of her perceptual and background beliefs, the agent can claim to have mistaken a piece of wood for a snake, for example,

¹⁸ For a review of the theories of valence, see Prinz (2004), Ch. 7. Prinz takes valence to be a different determinate experience in each feeling. On valence as determined by overall value, from a consumer semantics viewpoint, rather than as an experience of pleasure/displeasure, see Carruthers (2011), pp. 127–130. This view, however, does not build on the nonconceptual information being felt, but rather on its being represented “in an abstract and amodal way”, which, nevertheless, is motivating.

¹⁹ These expressive representations do not require a system to have the capacity to form propositional representations. They are close to what Strawson called “a featural representational system”, allowing an animal to navigate with no propositional thinking (1959). On the comparison between the two representational modes, see Proust (2013). The question of the penetrability of feelings by propositional thought is explored below, in section 5.

and to make explicit that there is no reason to be afraid of a piece of wood.

Our analysis of FS helps us to clarify why “feeling one’s keys in one’s pocket” does not belong to reactive feelings. Recognizing through touch the object in one’s pocket as being one’s keys, or merely having a proprioceptive experience in fact caused by one’s keys, are two ways of perceiving one’s keys, involving respectively cognitive and sensory proprioception. But neither needs as such to involve an affordance of a given intensity and valence. In contrast, let us suppose that the perceiver believes wrongly that she has forgotten her home keys, which are in some distant location, and will not be able to get back home. Feeling her keys in her pocket immediately triggers a positive affordance, opening up the field of possible actions.

4 Questions and objections

The present proposal raises a number of additional questions and objections. Let us start with the most radical objection.

4.1 Are feelings representations?

Granting that feelings, affective or not, can be pure “physical effects of objects on the nerves”, in [William James’](#) terms (1890, vol. 2, p. 458), they do not need to have any genuine representational value. James invites us to take the case either of purely somatic feelings or of objectless emotions when they are generated by a pathological condition—such as the precordial catch syndrome (PCS) which is a feeling of pain in the chest that usually goes away without treatment, but can lead the victim to think he or she is suffering a heart attack. In this case, the emotional experience of dread, [James](#) says, is “nothing but the feeling of a bodily state, and it has a purely bodily cause” (1890, vol. 2, p. 459). From this, one might conclude that a feeling is a merely peripheral phenomenon: it does not have a function to represent, nor does it express anything in particular. What can be said, in response, is, first, that feelings have a crucial evaluative function, which they perform thanks to their expressive structure. In PCS, the pa-

tient’s experience of dread has valence and intensity, expressed through sudden breathlessness, chest constriction, blurred vision, tingling sensations in the skin, an elevated heart beat, and a disposition to crouch. These feelings are not only a matter of sensory “peripheral” experience: they are also used by the patient to collect her existing Bayesian correlations, and to monitor with their help the present affordance expressed. A second illustration of the representational nature of feelings is that they can arise in the absence of the sensory basis they seem to have. For example, illusory feelings of being touched—a reactive somatosensory feeling about a change occurring on one’s body surface—can be created by manipulating the coherence of the intermodal inputs from vision, touch, and proprioception. In the so-called “rubber-hand illusion”, participants feel that their hand is being touched with a paint brush, when in fact it is an artificial hand, not theirs, that they see being touched. They also, after a while, “feel as if their (real) hand is turning ‘rubbery’” (see [Botvinick & Cohen 1998](#)). This experiment is evidence that feelings are informational states, which monitor inputs, and, in extreme cases like this, cause the brain to try to reconcile contradictory multimodal input. In the proposed interpretation, however, seeing one’s hand being touched is a reactive feeling, while actively touching an object generates a percept—which plays quite a different role in cognition.

4.2 What does “resonating to an affordance” mean?

Second, speaking of “subjective resonance” to an affordance (see the discussion of how a feeling “resonates” to an affordance in section 3 above) may look improperly metaphorical.²⁰ This is meant, however, to mark the difference between feeling and perceiving. While percepts allow recognition and identification of external objects and properties, feelings express specific affordances in a perceived, imagined, or remembered situation. For example, one can feel cold right now, or simulate being cold when

²⁰ In a similar vein, [William James](#) writes that, in emotions, “the whole organism is a sounding board” (1890, vol. 2, p. 450).

planning a polar trip; one can remember how angry, or bored one was in a given episode and context. Feelings give agents prompt access to the relevant features of a new situation through sensed changes in their experience. Importantly, resonance is also an apt term for empathy, i.e., for the propagation of feelings from an agent to an onlooker, based on expressive behavior (Decety & Meyer 2008; Dezechache et al. 2013). Brain imagery suggests that the perception of pain in another individual largely overlaps with the regions activated when experiencing pain oneself (Jackson et al. 2005). Such empathy, in the present proposal, exemplifies how a feeling structure can be communicated through a set of congruent behavioral cues associated with a given affordance (here a painful stimulus), with a valence and intensity that are bodily conveyed.

4.3 Non-conceptual content as a common feature of feelings and percepts

Third, one might object that a common feature of feelings and percepts is that they include non-conceptual contents. This is true; but notice the difference between the two types of non-conceptual content: while non-conceptual ingredients in perception are related to objective, external contrastive cues such as shapes, edges, colors, volumes, and auditory patterns, which can be static or dynamic, but are always purely descriptive, non-conceptual contents in feelings only include evaluative states, which combine the general type of the affordance, its valence, its intensity, the proper action program, where all constituents are “bodily marked”, i.e., expressed through somatic markers. Therefore we cannot say that feelings “perceive” affordances, for this would suppose either that feelings have direct sensory access to the world—which they don’t, for they extract their inputs from sensory perception—or that they have direct sensory access to the body, which they don’t have either—feelings are the subjective counterpart of bodily changes. Therefore we cannot say that agents “perceive affordances” when they experience a feeling, for this would suppose either that feelings have a direct sens-

ory access to the world, which they don’t, for they extract their inputs from sensory perception, or that they have direct sensory access to the body, which they don’t have either. Feelings are the subjective counterpart of bodily changes.

Neuroscientific research about the role of emotion in perception offers evidence in favor of this view. An affordance is made immediately salient by the system’s ability to sensitively react to a (half-)perceived element in a given known context.²¹ We speak of “half-perception” on the basis of what is known about the timing of object perception. Affordance predictions are made only milliseconds after visual sensations register on the retina, i.e., before the categorisation of perceived objects is completed (Barrett & Bar 2009). The orbitofrontal cortex (OFC; involved in emotion and reward in decision making, thanks to projections from the thalamus) is able to extract an affordance in the first 80ms of the visual process, merely on the basis of low spatial frequency and magnocellular visual input (Lamme & Roelfsema 2000). What happens to perceptual access when a perceiver cannot extract affordances? Barrett & Bar (2009) have shown that the lack of emotional reactivity in early perception impairs object categorization. A patient who accidentally lost his visual ability when three years old received in adulthood a corneal transplant. In spite of his recovered ability to extract visual information from the world, this perceiver had trouble categorizing what he saw. The authors’ suggestion is that reconstituting the internal affective context associated with past exposures to an object (which was lacking in this particular case) is “one component of the prediction that helps a person see the object in the first place” (Barrett & Bar 2009, p. 1325).

In summary: the medial OFC uses early low-level visual output to match the affordance associated with it in past experience of the object: somatic markers are thereby activated, and the appropriate action is prepared. A FS enables an object to be more swiftly categorized

²¹ For a defence of this view in terms of situated cognition, see Griffiths & Scarantino (2009). The authors emphasise the environmental scaffolding that makes possible affordance detection in emoters.

at higher perceptual levels. This evidence suggests that affordances are extracted from perception, but that feelings are not themselves perceived.²² On the contrary, they offer a separate kind of feedback to cognitive perceptual processes.

4.4 Respective role of somatic markers and formal content

Let us turn now to one of the most central questions that our proposal raises. How does it explain the respective roles, in expressive intentional content, of somatic markers, on the one hand, and of the represented formal objects on the other? Cognitive theorists take emotions to represent both salient aspects of the agents' own bodily changes and an evaluative belief about an external fact, with, possibly, a causal relation between this fact and the experienced bodily change (see [Gordon 1987](#); [Tye 2008](#) and [Solomon 2007](#)). For example, when perceiving a bear in the near vicinity, one's experience is taken to be about a complex of subjective bodily impressions (a pounding heart, trembling legs, etc.) and about the perception of a bear as being the cause of these changes. Such a construal of the intentional content of feelings only makes sense within a propositional mode of thought. Can our expressive mode reflect or approximate the information contained in this complex causal structure?

Clearly, FS does not *explicitly* convey a causal relation between situation, somatic markers and subjective feeling. It carries this causal relation implicitly, however, as a consequence of the control architecture that produces feelings. In an emotional control loop, a perceived affordance causes (rather than being represented as causing) its expressive evaluation through its specialized sensory feedback. Emotional awareness expresses this functional relation. An external event (made accessible through a perceived affordance, as detailed above) is immediately followed by subjectively experienced somatic cues of a given intensity and valence. In functional terms, this sequence makes sense in

the following way. When an associated forward model has been selected (often automatically, on the basis of an environmental, somatic, or cognitive affordance), the associated sensory cues (the somatic markers in this particular episode) are automatically activated in order to monitor how this affordance is to be processed and reacted to. As has been shown elsewhere, monitoring implicitly carries information about the command (or the affordance) that is being monitored (see [Proust 2013](#)). This explanation is particularly detailed and convincing in the case of motor representations of action; the feelings of agency that result from the comparators associated with a given feedforward model express (among others) whether the emoter is, or is not, the author of the action currently attended to (see [Wolpert et al. 2001](#) and [Pacherie 2008](#)). The present proposal generalizes the functional significance of feelings throughout their diverse types (reviewed in section 2). As the outcome of sensory comparators, feelings always carry a structured information set about the type of affordance they contribute to regulating, about its amount, and about which actions are appropriate. This information, in its own expressive mode, functionally approximates a causal relation that is, when propositionally expressed, represented as a relation between an internal state, an external cause, and a disposition to act.

In summary: Feelings do not gain their aboutness through a propositional thought where the contrast between object and property is semantically marked; they gain their functional (rather than propositional) aboutness (f-aboutness) through the respective roles, in adaptive control, of the selection of an affordance-dependent control model and of the markers that allow comparisons of valence and intensity to be expressed.

4.5 The attribution problem

This account, however, fails to explain observed variability in the production of feelings and the interpretation of what feelings are “about”. There are cases where agents misattribute their sadness, their anger, or their happiness to an

²² When we say that a feeling is felt, “felt” is not intended to mean “perceived”, but, rather, “entertained”.

event that is either not real, or that actually played no role in feeling production. How can such a misattribution be explained on the present proposal? Our first attempt to address this question is based on the subjective grounding of affordances. “Feeling *f*” normally means that an affordance is sensed, expressed, and subjectively represented as present. This does not mean that the affordance has an objective counterpart. Thus a thirsty traveller can be delighted or relieved when subjected to a water mirage. It is no problem for this view, then, if an event does not have the action potential for a given affordance it is expressed as having.

A trickier problem for the proposal is that a person might feel an *f*-feeling while she thinks that she has a *g*-feeling. Is such a situation even possible? To deal with this question, we must first clarify what “transparency” means when applied to feelings. A mental state is transparent if, when it is activated, its intentional content is accessible to the subject who entertains it, while its vehicle properties are not. On the view defended above, feelings are transparent, because their somatic markers are felt in connection with a certain affordance, and because their valence and intensity directly influence the emoter’s motivation to act in a given way. Such transparency, however, does not need to entail the subject’s ability to verbally report the content of her feeling. First, as seen above, a feeling can be felt by a nonhuman or by a child, both of whom lack the requisite verbal and conceptual capacities. Second, even an agent endowed with language can express through somatic markers a feeling with a distinctive FS content while failing to accurately report, in conceptual terms, what her feeling is “about”. We saw that [aboutness], i.e., reference to an independent event or object, is not a concept that belongs to FS. When subjects try to infer [aboutness] from their experience, their propositional system of representation (PS) is solicited. Because the latter has an analytic rather than an evaluative function, additional constraints step in. While nonconceptual, intensive (analog) and value considerations and norms regulate FS, conceptual, digital, and in-

strumental considerations and norms regulate PS.²³

Hence, when having to report about her feelings, a subject needs to translate one mode of representation into another, with no guarantee that this translation will not enrich or modify FS intentional content. First, she may no longer have access to the rich diversity of her FS experience, because her attention is no longer directed toward the relevant contextually-activated affordance. Second, she has to monitor other goals and their corresponding (social, instrumental, or epistemic) norms. For example, she needs to present her feelings to herself and to others in a socially acceptable way, and to try to justify them rationally. This in turn will depend on her existing background beliefs, on her self-concept, on her capacity for making self-attributions of this particular kind, and on her willingness to perform this kind of introspective report. A number of experiments and novels have documented the wide gap between people’s feeling experiences and the verbal report they provide, or the reasons they offer, for having this or that feeling. These considerations suggest, then, that the issue of transparency cannot be adjudicated independently of one’s viewpoint about mental architecture.²⁴ According to the present proposal, an affordance is *first* subjectively recognized through the resonance it produces—through its specific feeling, rather than through a concept-based interpretation.

Let us now return to our earlier question. Can a person actually feel an *f*-like feeling, and mistake this *f*-feeling for a *g*-feeling? According

²³ About the nature and role of nonconceptual norms, see Proust (2013).

²⁴ An alternative proposal by Carruthers (2011) sees as a condition of transparency of an affective feeling, rather, that the corresponding appraisal include the detection of the details of the associated non-conceptual somatic markers, which makes the recognition of a specific emotion possible, as well as its subsequent global broadcast—hence making this information available to the mindreading system. This analytic view of feelings, however, makes it utterly mysterious how a given pattern of autonomic measures is ever recognized, among thousands of similar patterns, as distinctive of an emotion. On the present view, a feeling is produced within a given forward model, which automatically activates the comparator for this affordance. Transparency, then, is effective only when a given forward model is activated, and does not need to transfer to a verbal modality. This seems to be recognized in part by Peter Carruthers, when he concludes that “we can have transparent access to the strength of only our occurrent *context-bound* affective attitudes” (2011, p. 146).

to the present account, this situation would presuppose that an *f*-feeling, as it occurs in the expressive mode, is misdescribed in a verbal report as a *g*-feeling, to finally be genuinely felt to be *g*. On this view, a change in representational form would not only make it possible to reinterpret the initial experience in terms of a different one, but also to feel differently. To see whether this case is plausible, it is worth discussing Schachter and Singer's (1962) adrenaline experiment.

5 Do beliefs influence affective report?

Schachter and Singer's famous adrenaline study aimed to collect evidence in favor of a two-factor theory of emotion, according to which a changed state of arousal leads agents to form feelings with a given valence that depends only on the epistemic/motivational context. Participants' arousal was manipulated by injecting them, under pretext, with adrenaline or a placebo. Only a subgroup of the adrenaline participants were informed that they had received a drug that would modify their arousal level. Participants were subsequently invited to stay in a waiting room where a confederate was either pretending to be euphoric or angry. Participants' emotional responses, observed in their behavior and subsequent self-report, differed in the various conditions: those unaware of having been injected with adrenaline, and placed in the anger condition, felt angriest, followed by the placebo + anger subjects. The least angry were the adrenaline informed participants. In the euphoria condition, misinformed adrenaline participants were "somewhat" happier, adrenaline informed ones somewhat less happy (in the euphoria condition, the results failed to reach significance both for behavior and self-report).

Were Schachter and Singer successful in making the point that valence of a feeling is a matter of attribution of the source of an experienced arousal? Several powerful objections have been raised against this claim. Recall that subjects were asked to what degree they would describe themselves as happy or angry. A first problem is that the questionnaire *suggested* the relevant target categories of emotions, which is

disturbingly close to influencing participants' responses (see [Plutchik & Ax 1967](#) and [Gordon 1987](#), p. 100). Furthermore, as noted above, ex post-facto reflective labeling of one's emotion does not need to express one's original feelings. As shown by [Nisbett & Wilson \(1977\)](#), self-reporting is highly sensitive to rationalizations from context. A second problem, mentioned by the authors in the discussion, is that the subjects' verbal reports and emotional behavior failed to confirm expectations in the euphoric condition. A third methodological problem, also recognized by the authors, is that the student participants had their own independent reasons for feeling anger in passing this longish test, which predisposed them to feel anger. There are, however, more theoretical objections.

On Schachter and Singer's view, the core feeling of an emotion is an arousal change, which can be artificially induced by drugs. Valence is supposedly gained through contextual beliefs and motives. If this view is accepted, why should we expect that contextually relevant beliefs specify the feeling itself (e.g., the anger experience)? Participants may indeed have been led to believe that they were angry when they were actually merely aroused. This does not show, however, that they ever felt anything else than an arousal change ([Gordon 1987](#), pp. 100–101). Schachter and Singer may have only biased self-attributions and self-report toward target emotions. The behavioral changes that were observed and attributed to felt emotion, in addition, can be imputed to social influence, rather than to intrinsic changes.

A final worry is that inducing in a participant a somatic marker normally associated with a given feeling (e. g., increased heart rate), *and* providing the person with a context rationalizing this somatic change, does not amount to an ecological way of producing a feeling. A cognitivist theorist of emotion will insist that the mere association between a physiological cue of the feeling *f* and a context does not amount to the realization, by a participant, that she feels *f* *because* she is in such and such a context ([Gordon 1987](#), pp. 98–99).²⁵ As discussed in section

²⁵ As Gordon observes, "one will not experience fear unless one connects up that cognition with the arousal one feels. To do this re-

4, the expressive mode has a nonconceptual representation of this causal connection. The architectural relation between feelings and affordances explains why subjects experience a systematic connection between their feeling and what it is “about”, much in the same way that an agent experiences a systematic connection between an intention to move and the goal that is aimed at—that is, without needing to represent conceptually the causal connection between the two. Nothing prevents the emoter, however, from forming a secondary conceptual representation of the emotional experience she has had, and reappraising the context on the basis of her background beliefs. As a consequence of this concept-based reappraisal, the agent may either discount the relevance of her initial feeling (as in the fear-of-snake case), or redescribe it in the richer terms that she now has available (as was done, presumably, by the Schachter and Singer participants).

Taken together, these objections have led most theorists to reject Schachter and Singer’s two-factor theory of emotion, and to look for alternative accounts of the role of inferences in self-attribution of feelings. It is interesting to see, however, that a two-factor theory has also been applied to the case of M-feelings.

6 Are metacognitive feelings sensitive to beliefs and inferences?

What are metacognitive (also called *noetic*, or *epistemic*) feelings? Juxtaposing [being metacognitive] and [being a feeling] sounds, at least prima facie, dangerously close to an oxymoron. When Descartes, Locke, and other 17th-century philosophers explored the properties of ideas as being “clear”, “distinct”, “evident”, and “certain” they certainly never took them to be feelings. These notions were taken, rather, to be objective representational properties that the mind, unaided by imagination, is able to detect. David Hume, in contrast, observed in his *Treat-*

ise that “the vivacity of the idea gives pleasure”, and that “its certainty prevents uneasiness by fixing one particular idea in the mind, and keeping it from wavering in the mind of its objects” (Hume 1739/40, 2007, p. 289). Thus Hume was glad to accept that epistemic feelings exist, and that they vary in their vivacity and in their pleasantness, i.e., in their intensity and in their valence. Following Hume’s lead, let us test how our analysis of FS above fares with the case of noetic feelings. Here, again, is our proposal about the general structure of feelings.

- FS Affordance_a [Place_a=here],
[Time_a=now/soon], [Valence_a=+/-], [Intensity_{a=n}(comparatively specified on a scale 0 to 1)], [motivation to act_a of degree_d according to action program_a].

What is specific to noetic feelings is that the affordances to which the system resonates are “informational” or “metacognitive” rather than environmental. Hence, the affordance does not relate to the external environment (the “here” slot is often irrelevant, except for perceptual affordances, or place-dependent metacognitive affordances, such as concentrating in a noisy spot). Although a cognitive action does not, in general, consist in physical moves towards or away from an affordance, similar decisions are motivated or inhibited in the domain of mental agency: a high retrieval affordance motivates pursuing the memory search, a low one to quit, etc. Hence our FS analysis also applies to noetic feelings.

As already emphasized, the affordances expressed in feelings do not need to be construed conceptually in order to be detected and assessed through their associated somatic markers. A conceptual construal, however, is suggested by the names given, in the literature and in ordinary language, to M-feelings. The term “feeling of knowing” (in response, for example, to the question: “what is the capital of Australia?”) implicitly presupposes that the emoter has access to the concept of knowledge. Expressing her feeling verbally, indeed, an emoter might say: “I feel that I know the response to this question”. In this sentence, she indeed refers

quires, according to him, a second cognition: a recognition or belief that is one’s being (or taking oneself to be) in a situation of danger that is causing the arousal one feels. This “cognitivist” objection is correct when targeting S and S’s theory, who also defend a cognitivist view of feelings. The present view, however, proposes a non-doxastic account of feelings, and is thus immune to this objection.”

to her disposition to retrieve knowledge and, hence, metarepresents her knowledge disposition.²⁶ The affordance theory of noetic feelings suggests a different picture. When trying to remember a proper name, a *feeling of knowing* is a specific experience of having the ability to detect the target, and of predicting its imminent recall. It can be associated with a feeling of tension (Koriat & Levy-Sadot 1999, p. 486). This experience is associated, then, with a graded, intuitive, and affect-like appraisal of a [remembering] affordance. Rhesus monkeys working in experimental labs in comparative psychology show that they can assess their memory affordances (see Beran et al. 2012, Chapter 1).²⁷ What kind of feedback, then, do monkeys use? A surprising and substantive fact about metacognitive control, first revealed through the pioneering research of Asher Koriat, is that the comparator generating metacognitive feelings (such as a feeling of knowing in a memory task, or a feeling of clearly discriminating in a discrimination task) has no access to the semantic contents stored in memory or made available through perception. In Koriat's words, M-feelings "are mediated by the implicit application of non-analytic heuristics, relying on a variety of cues." These cues "pertain to global, structural aspects of the processing of information", such as ease of processing, time devoted to a task, familiarity, and accessibility (Koriat 2000; Koriat & Levy-Sadot 1999).²⁸ Therefore, contrary to what epistemologists have always believed, the most common type of epistemic appraisal is not directly based on the content of the thoughts to be evaluated, but on the properties of the underlying informational process.

Neuroscientific research confirms Koriat's claim. Implicit, associative cues are extracted by the working brain to select, in a cost-efficient

way, what there is to learn, to retrieve from memory, to extract from perception, or what is worth storing in memory. These are all to do with the dynamics of information processing: with its onset, with the comparative amount of activity in incompatible neural responses, and with the time needed to converge on a threshold value. Indeed, the neural activity recorded in rats' OFC when attempting to categorize olfactory stimuli was found to correlate with their predictive behavior (consisting in accepting or rejecting a task trial); similar patterns have been found in other species.²⁹

On the FS model, somatic markers have the function of expressing the intensity and valence of the noetic predictions generated from feedback at the neural level. As indicated in section 2, psychophysiological measures (electromyography) provide evidence for the existence of facial markers associated with feelings of fluency and of disfluency (Winkielman & Cacioppo 2001). Increased activity in the smile muscle, the zygomaticus major, produces feelings with a positive valence. A reduction of fluency is correlated with activity in the corrugator supercilii (involved in frowning), which suggests that this additional effort is felt as unpleasant. Intensity of positive or negative confidence, computed implicitly, is expressed by the corresponding intensity of the noetic feeling. A different somatic marker of memory appraisal is the TOT phenomenon. This often occurs when a search in memory for a specific word fails to retrieve that word within the usual time interval. The informational ingredients of FS are conveyed by the intensity of the activity in the tongue muscle, and by the affective quality of TOT. Taken together, these predict the likeli-

²⁹ See Kepecs et al. (2008). An interesting account of the predictive activity reflected in noetic feelings is that the dynamic activity in the neurons activated by a given task correlates with the so-called "accumulation of evidence" that is diagnostic of success or failure in that task. For example, in a perceptual discrimination task, where a target might be categorized as an *X* or as a *Y*, evidence for each alternative is accumulated in parallel, until the difference exceeds a threshold, which triggers the perceptual decision. The information that will generate a feeling consists, first, in the differential rate of accumulation of evidence for the two (or more) possible responses, and second, in stored information about the threshold value, computed from prior trials, which the rate of accumulation should reach in order to make a cognitive decision likely to be correct. For a discussion and review of the literature, see Fleming & Dolan (2012), and Proust (2013, pp. 99).

²⁶ Arango-Muñoz (2012) claims that feelings of forgetting and feelings of knowing are cases of "conceptual experiences". According to the present view, following the lead of Koriat and colleagues, M-feelings can overlap with judgments, and be redescribed in conceptual terms; they pertain, however, to different representational levels. There are no "conceptual experiences", except in the sense of experiencing the comparative fluency of concepts.

²⁷ As indicated above, rhesus monkeys are able, in a perceptual or memory task, to opt out of more or less challenging trials as a result of trial difficulty.

²⁸ As will transpire below, all these cues are, as far as we know, dimensions or effects of fluency, i.e., of ease of processing.

hood of successful retrieval. An implicit cue-based heuristic might thus explain why TOTs have the valid predictive value they do (Schwartz et al. 2000).

6.1 Two-factor theories of M-feelings

In our FS single-factor model, M-feelings have an intrinsic intensity and an intrinsic valence. Two-factor theories make a different claim, in ways analogous to Schachter and Singer's theory of aboutness in affects: M-feelings have an intrinsic arousal level, but their valence depends on the environment. Jacoby and his colleagues were the first to embrace a two-factor view about feelings of fluency. They manipulated participant's exposure to an item in order to show that enhanced fluency generates an illusory feeling of familiarity. Under conditions of divided attention, reading a list containing both famous and not famous names raised participants' disposition to wrongly judge as famous some names presented in a second list, merely because these names had already been read in the first list. Schachter and Singer's idea was that fluency is a generic feeling, that needs to be interpreted on the basis of goals and current cues, in order to deliver a qualitatively different specific feeling:

Inherent in the idea that the subjective experience of familiarity arises from an interpretation of cues is the notion that cues can be interpreted in a variety of ways. As noted above, if ease of identifying an item is obviously being manipulated by the experimenter, the resulting perceptual fluency does not give rise to a feeling of familiarity. Attributions are also affected by one's goals. In the context of attempts to remember, people may be more likely to interpret ease of generating an item or perceiving it as familiarity. In the context of other tasks, the same cues may be interpreted in other ways. (Kelley & Jacoby 1998, p. 129)

From their viewpoint, the fluency generated by a given name can, according to the task and the

information made consciously available to a participant, be experienced as a feeling of familiarity, or as a feeling of recognition of that name as "old" (i.e., presented in a former list). They conclude that a feeling of fluency (generated by a perceived name) will be experienced as a function of the alternative ways of interpreting this feeling, on the basis of the agent's goals and the additional cues available.³⁰

A similar two-factor theory has been defended in the (Whittlesea & Williams 2000; Whittlesea & Williams 2001) model of M-feelings. According to this model, feelings of familiarity result from the perception of a *nonspecific* discrepancy between the expected and the observed rate of processing of elements in a given context. Valence and the associated action guidance, on the other hand, are based on a conceptual interpretation of what this discrepancy means. For example, you find yourself waiting for the bus next to people you expect to be total strangers. Suddenly, you have an unexpectedly high fluency experience when looking at the face of someone you have already encountered several times—a clerk from the local grocery shop. This unexpectedly high rate of discrepancy-reduction determines an intense feeling of familiarity with a strong motivation to identify the familiar face (see Whittlesea & Williams 2001). Had you seen the clerk in the local grocery store instead, you would have merely had a feeling of recognition when seeing the clerk.

To summarize: the core idea in two-factor accounts is that participants have a primary feeling of fluency, which they interpret in more specific terms as a function of their goals and of the context as they consciously represent it to be. Thus, on this view, a feeling partly relies on background knowledge, and partly on a naïve theory concerning the relation between feelings and mental activity (Schwarz & Clore 2007). The naïve theory is as follows: feelings are about what one is doing, so this feeling must be about this event of trying to perceive, or this attempt at retrieving, etc.

³⁰ Jacoby & Whitehouse (1989) similarly argue that a feeling of fluency can be experienced as familiarity in a memory task, and as confidence in a problem-solving task.

As already observed above, a naïve-theory view is incompatible with monkeys' and young children's epistemic evaluations based on fluency. Our FS structure offers an alternative account: cues (associative heuristics) dictate how an affordance is detected, assessed, and exploited in a context, but these cues are not consciously available, and hence do not depend on a naïve theory of the task. The Jacoby and Whitehouse evidence is compatible with a procedural view of engagement in a task through automatic memory processes, and of the feelings of familiarity they generate. A comparator is always activated as a function of a subject having been highly trained in the corresponding first-level cognitive task. Monkeys and humans feel that a memorial or perceptual affordance is present because, if they need to assess whether, for example, an item was seen earlier, the associated comparator produces a feeling of a given intensity and valence indexing the remembering affordance. Thus, it is uncontroversial that a context-dependent factor determines both the task to be performed and the reactive metacognitive feeling about this task.

It does not follow from the context-dependence of a cognitive task, however, that a concept-based interpretation will affect the experienced feeling itself, as maintained by the two-factor theorist. A cue-based, non-analytic heuristic is not inferential in the interpretive, first-person sense. Regrettably, the word "inference" has been loosely used in affective and in metacognitive studies, to refer both to "automatic, non-analytic, largely unconscious and fast associative processes" (Nussinson & Koriat 2008) and to conscious reasoning and theory-building (Schwarz & Clore 2007). These two types of processes (respectively called "automatic" and "controlled"), are now held by many authors to operate independently.³¹ While unconscious heuristics rely on implicit associations between cues, inferences comprise deductions from premises to conclusions. Looking back at Jacoby and Kelley's point above, we see that

the authors are referring to unconscious cues being recruited for a task: they are thus referring to unconscious associative heuristics rather than to explicit concept-based reasoning. The memory interactions they are exploring, however, typically involve both automatic and controlled processes, which is a source of confusion. As Jacoby and Kelley are eager to show, implicit associations and explicit reasoning lead to different, incompatible predictions. As a result, the evidence they present shows how automatically-generated feelings can be theorized about in controlled processes. It does not demonstrate, however, that feelings depend upon theorization. A theory of the task, in contrast with automatically generated feelings, offers reasons to attribute to oneself beliefs and motivations to act, and, possibly, to reject the relevance of feelings for any particular task.

Our proposal, then, has several advantages over inferential or theory-based accounts of fluency. First, it explains why a feeling of fluency can be experienced, and why it can motivate agents' metacognitive responses in species or individuals with no concept-based attributive capacity (i.e., with no capacity for mindreading). Second, our proposal accounts for the difference between a type of M-feeling (a feeling of fluency) and the various ways in which it is experienced across cognitive tasks. Granting that comparative ease of processing can *always* be computed, and can be used as a reliable indicator of the likelihood of success across a wide range of cognitive activities, it is not surprising that there is a type of feeling based upon it. Fluency can be perceptual, memorial ("retrieval fluency"), or conceptual. It can be used in predictive or retrospective evaluations. If agents are asked to determine which statements are likely to be true or false (presumably a question that only—but not all—humans can understand), felt perceptual fluency will induce a "truth effect". Agents will evaluate a statement as more likely to be true than another merely because it is easier to read.³² If agents are asked

³¹ For a defence of the distinction see Jacoby & Brooks (1984), Koriat & Levy-Sadot (1999), Recanati (2002) and Smith & DeCoster (1999). Koriat & Levy-Sadot (1999) both emphasize the distinction and use the term "inference" in both cases.

³² There is abundant evidence, however, that M-feelings uncritically guide epistemic decision (i.e., are unopposed by concept-based processes) mostly when the cognitive task is unimportant, when cognitive resources are limited (under time pressure or divided attention), and when agents are in a good mood (Nussinson & Koriat 2008; Schwarz 2004).

to detect faces of known people (or of stimuli previously shown), felt fluency will generate a sense of familiarity, which motivates agents to try to identify the target. If people are asked to assess the frequency of a given phenomenon, felt retrieval fluency—that is, what comes immediately to mind—will be used to judge what is more frequent. Felt fluency will also have effects outside of metacognition: if agents are asked which particular face, landscape, or picture they prefer, felt fluency will influence their decision. Several affordances, then, may be associated with the same globally expressive *type* of feeling (constructed as the set of feelings with the same type of facial markers for ease of processing, for example). The notion of type of feeling is a technical term, which is useful to distinguish the diverse ways in which fluency is used by the brain. But a type of feeling is never experienced; only tokens of the type are. Tokens of feelings of the same type will differ in the specific affordances that are detected, and in the tendencies to act that the feeling motivates. As a consequence, one cannot say that feelings of fluency “feel the same” to an emoter: fluency experienced in an FOK and in an FOR, for example, apply to different segments of processing, assess different things, and motivate a different action program. You may first have an FOK after a question is addressed to you, and then fail to have the associated FOR after having come up with a response. These differences have nothing to do with an interpretation: they are constitutive of what sensitivity to a given affordance amounts to. Take the case of feelings of familiarity. As summarized [above](#), Whittlesea and Williams claim that fluency is the core of the experience, while familiarity is a conceptual interpretation of this core feeling. It is more economical, however, to suppose that familiarity is a different feeling within the general fluency type, and that it is associated with a different affordance.

In summary: engaging in a particular cognitive task (e.g., trying to remember, evaluating retrieval, assessing frequency) does not need, per se, to involve a naïve theory of the task. It only requires having a salient affordance, and an implicit heuristic for metacognitive predictions in that task.

6.2 Incidental versus integral feelings

Our proposal also allows us to address in affective terms the issue of incidental versus integral feelings, which, in the literature, is invariably framed in inferential terms (with all the ambiguity relating to this expression). Metacognitive feelings are called “incidental” when they are not based on valid cues for the cognitive task at hand, and hence, have no predictive value. They are called “integral” when they actually carry information about cognitive outcome. Granting the universal role of fluency in metacognition, how do people know when a feeling of fluency is relevant to a given task, and which sequence of their cognitive activity needs to be monitored? A frequent answer, in the literature, is that agents believe that fluency applies by default to the present domain of judgment. When, however, agents are led to believe that a feeling of fluency is purely incidental to the task at hand, they will discount it in their decision, on the basis of a theory of the domain of interest (see [Schwarz & Clore 2007](#) and [Whittlesea & Williams 2000, 2001](#)). Let us suppose, in what we shall call case (a), that an agent is explicitly told that a given cue, such as the ease of reading a given sentence, is irrelevant to a given task—such as assessing the truth value of the written statement. Or, alternatively, let us suppose—case (b)—that the agent discovers by himself that there is a connection, but with reverse relevance. Perhaps he finds that badly written sentences, involving added processing effort—in a given context—are likely to be true (see [Unkelbach 2007](#) and [Unkelbach & Greifeneder 2013](#)). A popular account of these cases is that people will infer respectively, for (a): that the feeling of fluent reading they have had *is not about* the target task, which entails that reading fluency does not predict truth, or, for (b): that what predicts the truth of a written utterance, in this particular context, is disfluent reading (see [Schwarz & Clore 2007](#), p. 394).

According to this two-factor account, M-feelings are cognitively penetrable. They can be suppressed at will, on the basis of a reinterpretation of their being experienced, or can even be

used to predict falsity instead of truth.³³ On the account proposed here, in contrast, M-feelings are never cognitively penetrable. Why, then, do subjects stop trusting their feeling of fluency? Our answer is the following. In the first type of case, subjects do not allow their feelings of fluency to guide their decision because they have received verbal instructions to this effect. In the second type of case, subjects no longer use their feelings of fluency to form an epistemic decision in the proposed task, because they have learned, over time, that these feelings do not predict truth in this task.

In case (a), then, subjects are confronted with a different task. They are no longer asked to express their confidence in the truth of a given sentence (an intuitive, associative task); they are asked to assess the truth of sentences by taking into account the fact that their feelings of fluency are irrelevant. This new task requires the participants to form appraisals based on analytic reasoning. Feelings no longer drive their evaluation and epistemic decision.

In case (b), where bad writing is associated with likely truth, no “theory of the task” needs to be formed, on top of the first-order task, which consists in judging whether a written statement is true or not. A mere change in cue validity can produce, over time, a change in associative heuristics, and, hence, in feelings and in decisions to act. For example, just as our thirsty traveller will eventually learn not to trust an apparent “drinking affordance”, an agent will learn, in certain recurrent contexts, not to trust an apparent “fluency affordance”. Obviously, cue validity can, in humans, be conveyed verbally; this will considerably abridge the revision process of the associated program of action. We then return to case (a): participants will be able to immediately discount an apparently valid cue, to turn to analytic appraisals, and to refrain from acting on their fluent feeling (which, however, is still there). Cue validity, however, can be learnt implicitly as

well, which weakens the case for a theory-laden view of feelings.

These observations suggest that feeling-based and analytic appraisal, as hypothesized in this proposal, “tap separate databases representing knowledge in different formats.”³⁴ A feeling of fluency, as a result, can survive being discounted in decision-making. Another finding points in the same direction. There is evidence that, even when an M-feeling has been explicitly discounted (i.e., shown to agents to unduly bias their epistemic assessment), the initial feeling remains unaffected, and is able to promote further epistemic decisions. In Nussinson & Koriat’s (2008) study, agents exposed to unsolved anagrams and to anagrams accompanied by their solution, were asked to rate the difficulty of these anagrams for naïve participants with no prior access to the solution. The participants’ ratings were influenced by the differential fluency that the anagrams presented for them: the higher fluency of solved anagrams biased their attributions of difficulty. After being informed of the contaminating effect of knowing the solutions, the participants were invited to correct their attributions by re-rating the difficulty of the anagrams, which they did. However, the participants were subjected to a subsequent test, where, under time pressure, they had to predict which of two anagrams would be harder for others to solve. These other-attributions of difficulty presented, again, the same bias for known anagrams. Being under pressure allowed participants’ M-feelings to guide decision. The verbal instruction could shift their controlled responses when re-rating the anagrams, but did not lead the participants to recompute them, as should have been the case if feelings are cognitively penetrable.

In summary: what participants learned (that solved anagrams only *look* easier to process) did not influence what they felt later (higher fluency is diagnostic of ease of solving).

7 Are all feelings affective?

It is often noticed that a phenomenological contrast seems to exist between feelings—that is, they are not equally emotional. Are not M-feel-

³³ This two-factor account is endorsed by Unkelbach (2007): “the feeling resulting from the discrepancy is non specific, and the discrepancy triggers a search for an explanation [...]. The experienced variations are not attributed to prior exposure, resulting in a feeling of familiarity, but to some other quality of the statement, namely, that a statement is true.”

³⁴ A quote from Smith & DeCoster (1999), p. 329, who offer a strong defence of this view.

ings in general as “cold” as the proprioceptive feeling that my right arm is being extended? Or can they also be “hot”—that is, involve valence, i.e., be pleasant or unpleasant? Our proposal of a common expressive evaluative format suggests that all the feelings vary in affect in roughly the same way, because they all include valence in their informational structure. [Stepper & Strack \(1993\)](#), however, have emphasized that epistemic feelings are “cold”. Feelings like effort, familiarity, surprise, or feeling of knowing “have no fixed valence”, in the sense that they don’t feel particularly good or bad. Linguistic research on the emotional lexicon is invoked as congruent evidence: for words referring to readiness, success, and a desire to deal with new information (like “alert” “confused”), i.e., terms expressing metacognition, affects are not “focal”, which implies that they are not centrally emotional ([Ortony et al. 1987](#)).

There is abundant evidence, however, that feelings of fluency increase perceivers’ liking of the objects perceived. Familiar items (other things being equal) are found to be more pleasant than new ones. An initially neutral stimulus is felt to be pleasant after repeated exposure. This “exposure effect”, first demonstrated by Zajonc, has been attributed to increased perceptual fluency ([Zajonc 1968](#)). This affective effect of fluency has since been found to apply to any dimension of a perceptual input. The sense of beauty in a symmetrical face or in a landscape, or the pleasure felt in contemplating a picture seem to be inherent to the feeling of fluency generated in the perception. As noted above, psychophysiological measures in the facial muscles provide additional evidence for the affective character of the feeling of fluency ([Reber et al. 2004](#); [Winkielman & Cacioppo 2001](#); for a review see [Oppenheimer 2008](#)).

An interesting, untested, speculation intended to explain the presence of cold and hot versions of feelings is that valence, although never fully absent from monitoring, is modulated by dynamic aspects of the task under evaluation ([Carver & Scheier 1990](#); [Carver & Scheier 2001](#)). On this view, affective feelings can appear in physical and cognitive action, and probably also in somatosensory experience,

when certain dynamic conditions for affective reactions are present. But what are these conditions?

Let us first examine an area where these dynamic conditions seem to have a minimal role. This is the area of first-order motor control (including the initiation of an action, the monitoring of its development, and of goal completion). As with any other form of control, motor control involves specialized feelings, in the above sense of subjective experiences with a distinctive embodied phenomenal quality (see [Pacherie 2008](#)). At first glance, these feelings do not typically seem to be affective.³⁵ Why is this so? According to Carver and Scheier, this can be explained by the dynamics of a monitored activity that generates feelings. Affective feelings are part of a second-order type of feedback, having, in their terms, “the meta-monitoring function” of “checking on how well the action loop is doing at reducing the behavioral discrepancy that the action loop is monitoring”. This meta-loop, then, monitors a particular aspect of one’s progress in relation to one’s distal goal: it represents “*the rate of discrepancy reduction in the behavioral (monitoring) system over time*”. This dynamic representation is what a feeling is equipped to offer: the intensity and quality of a positive, or a negative, feeling express how far above, or how far below, the observed *rate* of discrepancy reduction is, with respect to some reference value. One consequence of this view, if it turns out to be experimentally validated, is fascinating and deep: affect in action does not depend merely on the amount of discrepancy being reduced. An agent may be an inexperienced performer in a task; if the velocity of her progress to the goal is higher than expected, she will feel more confident, and have retrospectively more positive feelings when reaching her goal than a competent performer whose progress to the goal is as steady as predicted.

There is a second type of affect, according to Carver and Scheier, that the dynamics of prediction can generate. Acceleration is the rate of change of velocity. Feelings express such acceleration when the rate of discrepancy reduc-

³⁵ Even in this domain, however, an error signal, when conscious, is associated with an unpleasant feeling.

tion *increases* beyond expectancy—a sense of exhilaration then occurs. Lucky athletes, who break several records within days, experience this. Symmetrical feelings of sinking, or despair, arise when the rate of discrepancy reduction *decelerates* unexpectedly and falls below the expected threshold more quickly than anticipated. In summary, cold motor feelings are generated when one is routinely acting on the world, when things develop as expected, except for small motor adjustments. Hot action feelings are generated when action monitoring involves unexpected dynamics of reduction or increment of likely success or failure.

How does this theory apply to M-feelings? A similar contrast may exist in M-feelings. Carver and Scheier's model allows us to predict that M-feelings can have colder and warmer varieties, depending on the dynamics of the discrepancy reduction that they express. As seen above, there are two varieties of M-feelings, distinguished by their function. Some, like FOKs, have a predictive function. Others, like FORs, perform retrospective evaluation. Neuroscientists explain these feelings through the rate of the accumulation of evidence, measured through the comparative activity of the neural assemblies involved in cognitive decision. (This rate of accumulation has to be compared with a stored standard in order to produce a reliable feeling of confidence.) From this widely accepted model, it follows that the rate of reduction of discrepancy toward a confidence threshold is automatically computed, and plausibly expressed through somatic markers that themselves have a varying intensity.

If this reasoning is correct, then although all M-feelings do not often have a definite “hot” quality comparable to fear and love, they always have a valence, according to whether they predict an agent's progress towards or away from her cognitive goal. To find more intense M-feelings, however, one needs to look at the dynamics of *meta-monitoring*, which is when an agent expects a given rate of reduction of the discrepancies toward her cognitive goal, and either observes a rate that is well above the expected rate or well below it. In these cases, the sense of confidence that the positively surprised agent

experiences is modulated by an intense, highly motivating affect of joy and renewed passion for the associated cognitive activity; while the uncertainty of the negatively surprised agent is associated with an intense, highly demotivating affect of discouragement, or loss of interest. Note how crucial an intense feeling of this kind can be, especially with regard to future motivation. It can precipitate in children a passion for learning; or it can lead them to reject an activity, or even a whole group of similar activities, because of the threatening affect associated with the activity, often combined with a still more threatening social affect (the sense of being an inferior, incompetent performer, or of being stupid). This kind of meta-monitoring cognitive affect, important as it is in predicting and fuelling epistemic motivation, is not easily observable in experimental settings, because it is elicited in middle or long-term forms of cognitive tasks, such as studying at school in a given grade, learning algebra, etc. This may in part explain why Stepper and Strack have failed to encounter it.

To summarize: noetic feelings, like all feelings, have an evaluative function. They are the output of a monitoring process, which expresses how likely it is that an agent's cognitive preferences or goals will be (or have been) fulfilled in a given task and context. They all have a valence, but their affective tonality is more intensely felt in special cases that arise when meta-monitoring makes “intensively new” affordances salient. The rate or the acceleration with which an observed initial discrepancy differs from a predicted standard value may either exceed the expected value, thereby producing positive feelings of confidence or feelings of knowing, or be insufficient to reach this value, producing negative feelings of uncertainty. The intensity of positive or negative affect in M-feelings thus depends on particularly unexpected properties of the underlying cognitive activity.

8 Conclusion

On the present proposal, “feelings” are not isolated sensory events. They are, rather, the

ingredients of a nonlinguistic expressive mode that allows organisms to evaluate and predict environmental changes and affordances. This expressive mode is of a relational, intensive kind that is not suitable for a predicative, concept-based representation of the world. As a consequence, feelings are not themselves judgments about the world or about one's own thoughts. They are not "about" anything in the objective, referring sense of the term. Feelings are able to approximate (in their own mode) the guidance offered by full-blown judgments, and hence can be re-described in conceptual terms when the latter are available to the emoter.

The importance of the duality between an expressive and a propositional system of representation has generally been overlooked. Even dual-processing theorists rarely appreciate that the two systems involved in cognitive evaluation and in reasoning have their own independent, although asymmetrical, role to play. A purely automatic, reactive type of evaluation is possible, and is present in nonhumans and young children. It is prone, however, to generating throughout life illusions of competence and reasoning errors. A conceptually-controlled type of evaluation, on the other hand, can partially inhibit the influence of the expressive system, but it still depends on the latter to weigh the impact of context on ability, and to assess the trade-off between ease of processing and informativeness—that is, relevance—that is crucial in communication and in problem solving.

A major practical consequence of the duality between the two target representational modes concerns pedagogy. Children cannot learn what they are *not* motivated to learn. Their motivation heavily depends on their subjective experience of what a school context affords them. Their feelings of confidence, i.e., the feedback from the cognitive tasks they engage in, have to be sufficiently positive and appropriately calibrated in order for them to form their own realistic and motivating cognitive goals. No amount of analytic reasoning can replace a positive experience when learning.

Acknowledgement

I am grateful to Dick Carter, Laurence Conty, Terry Eskenazy, Martin Fortier, Jonathan Frome, Thomas Metzinger and Jennifer Windt for their critical comments. Special thanks to Dick Carter for his linguistic advice, to Tony Marcel, whose critical questions about the commonality between M-feelings and affective feelings inspired the present article, and to Robert Gordon, for giving me access to some of his unpublished writings. This research has been supported by an ERC Senior Grant "Dividnorm" #269616, and by two institutional grants: ANR-11-LABX-0087 IEC and ANR-11-IDEX-0001-02 PSL.

References

- Arango-Muñoz, S. (2012). The nature of epistemic feelings. *Philosophical Psychology*, 27 (2), 193-211. [10.1080/09515089.2012.732002](https://doi.org/10.1080/09515089.2012.732002)
- Barrett, L. F. & Bar, M. (2009). See it with feeling: Affective predictions during object perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364 (1521), 1325-1334. [10.1098/rstb.2008.0312](https://doi.org/10.1098/rstb.2008.0312).
- Bechara, A., Damasio, H. & Damasio, A. R. (2000). Emotion, decision making and the orbitofrontal cortex. *Cerebral Cortex*, 10 (3), 295-307. [10.1093/cercor/10.3.295](https://doi.org/10.1093/cercor/10.3.295).
- Beran, M. J., Brandl, J., Perner, J. & Proust, J. (Eds.) (2012). *The foundations of metacognition*. Oxford, UK: Oxford University Press.
- Botvinick, M. & Cohen, J. (1998). Rubber hands 'feel' touch that eyes see. *Nature*, 391 (6669), 756-756. [10.1038/35784](https://doi.org/10.1038/35784).
- Brown, A. S. (1991). A review of the tip-of-the-tongue experience. *Psychological Bulletin*, 109 (2), 204-223. [10.1037/0033-2909.109.2.204](https://doi.org/10.1037/0033-2909.109.2.204)
- Carruthers, P. (2011). *The opacity of mind. An integrative theory of self-knowledge*. Oxford, UK: Oxford University Press.
- Carver, C. S. & Scheier, M. F. (1990). Origins and functions of positive and negative affect: A control-process view. *Psychological Review*, 97 (1), 19-35. [10.1037/0033-295X.97.1.19](https://doi.org/10.1037/0033-295X.97.1.19)
- (2001). *On the self-regulation of behavior*. Cambridge, UK: Cambridge University Press.
- Damasio, A. (1994). *Descartes' error: Emotion, reason and the human brain*. New York, NY: Putnam.
- (2003). *Looking for Spinoza: Joy, sorrow and the feeling brain*. Orlando, FL: Harvest Book Harcourt, Inc.
- Decety, J. & Meyer, M. (2008). From emotion resonance to empathic understanding: A social developmental neuroscience account. *Development and Psychopathology*, 20 (04), 1053-1080. [10.1017/S0954579408000503](https://doi.org/10.1017/S0954579408000503)
- Dezecache, G., Conty, L., Chadwick, M., Philip, L., Soussignan, R., Sperber, D. & Grèzes, J. (2013). Evidence for unintentional emotional contagion beyond dyads. *PLoS One*, 8 (6), e67371-e67371. [10.1371/journal.pone.0067371](https://doi.org/10.1371/journal.pone.0067371)
- Dolan, R. J. (2002). Emotion, cognition, and behavior. *Science*, 298 (5596), 1191-1194. [10.1126/science.1076358](https://doi.org/10.1126/science.1076358)
- Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, 6 (3-4), 45-60. [10.1080/02699939208411068](https://doi.org/10.1080/02699939208411068)
- Fleming, S. M. & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367 (1594), 1338-1349. [10.1098/rstb.2011.0417](https://doi.org/10.1098/rstb.2011.0417)
- Fodor, J. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.
- Frank, R. H. (1988). *Passions within reasons: The strategic role of the emotions*. New York, NY: Norton.
- Goldie, P. (2002). *The emotions: A philosophical exploration*. Oxford, UK: Oxford University Press.
- (2009). Getting feelings into emotional experience in the right way. *Emotion Review*, 1 (3), 232-239. [10.1177/1754073909103591](https://doi.org/10.1177/1754073909103591)
- Gordon, R. (1987). *The structure of emotions*. Cambridge, UK: Cambridge University Press.
- Griffiths, P. E. (1997). *What emotions really are*. Chicago, IL: The University of Chicago Press.
- (2004). Emotions as natural and normative kinds. *Philosophy of Science*, 71 (5), 901-911. [10.1086/425944](https://doi.org/10.1086/425944)
- Griffiths, P. E. & Scarantino, A. (2009). Emotions in the wild: The situated perspective on emotion. In M. Aydede & P. Robbins (Eds.) *The Cambridge handbook of situated cognition* (pp. 437-453). New York, NY: Cambridge University Press.
- Haggard, P., Clark, S. & Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature Neuroscience*, 5, 282-285. [10.1038/nn827](https://doi.org/10.1038/nn827)
- Hume, D. (2007). *Treatise of human nature*. Oxford, UK: Clarendon Press.
- Jackson, P. L., Meltzoff, A. N. & Decety, J. (2005). How do we perceive the pain of others? A window into the neural processes involved in empathy. *NeuroImage*, 24 (3), 771-779. [10.1016/j.neuroimage.2004.09.006](https://doi.org/10.1016/j.neuroimage.2004.09.006)
- Jacoby, L. L. & Brooks, L. R. (1984). Nonanalytic cognition: Memory, perception, and concept learning. *The Psychology of Learning and Motivation*, 18, 1-47. [10.1016/S0079-7421\(08\)60358-8](https://doi.org/10.1016/S0079-7421(08)60358-8)
- Jacoby, L. L. & Whitehouse, K. (1989). An illusion of memory: False recognition influenced by unconscious perception. *Journal of Experimental Psychology: General*, 118 (2), 126-135. [10.1037/0096-3445.118.2.126](https://doi.org/10.1037/0096-3445.118.2.126)
- James, W. (1884/1890). *The principles of psychology*. New York, NY: Dover.
- Kelley, C. M. & Jacoby, L. L. (1998). Subjective reports and process dissociation: Fluency, knowing, and feeling. *Acta Psychologica*, 98 (2), 127-140. [10.1016/S0001-6918\(97\)00039-5](https://doi.org/10.1016/S0001-6918(97)00039-5)

- Kenny, A. (1963). *Action, emotion and will*. London, UK: Routledge & Kegan Paul.
- Kepecs, A., Naoshige, U., Zariwala, H. & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, 455, 227-231. [10.1038/nature07200](#)
- Koriat, A. (2000). The feeling of knowing: Some metatheoretical implications for consciousness and control. *Consciousness and Cognition*, 9 (2), 149-171. [10.1006/ccog.2000.0433](#)
- Koriat, A. & Ackerman, R. (2010). Metacognition and mindreading: Judgments of learning for self and other during self-paced study. *Consciousness and Cognition*, 19 (1), 251-264. [10.1016/j.concog.2009.12.010](#)
- Koriat, A. & Levy-Sadot, R. (1999). Processes underlying metacognitive judgments: Information-based and experience-based monitoring of one's own knowledge. In S. Chaiken & Y. Trope (Eds.) *Dual-process theories in social psychology* (pp. 483-502). London, UK: The Guilford Press.
- (2001). The combined contributions of the cue-familiarity and accessibility heuristics to feelings of knowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27 (1), 34-34. [10.1037/0278-7393.27.1.34](#)
- Lamme, V. A. & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, 23 (11), 571-579. [10.1016/S0166-2236\(00\)01657-X](#)
- LeDoux, J. E. (1996). *The emotional brain*. New York, NY: Simon & Schuster.
- Logan, G. D. & Crump, M. J. (2010). Cognitive illusions of authorship reveal hierarchical error detection in skilled typists. *Science*, 330 (6004), 683-686. [10.1126/science.1190483](#)
- Mangan, B. (1993). Taking phenomenology seriously: The “fringe” and its implications for cognitive research. *Consciousness and Cognition*, 2 (2), 89-108. [10.1006/ccog.1993.1008](#)
- (2000). What feeling is the “feeling of knowing”? *Consciousness and Cognition*, 9 (4), 516-537. [10.1006/ccog.2000.0488](#)
- Murata, A., Moser, J. S. & Kitayama, S. (2013). Culture shapes electrocortical responses during emotion suppression. *Social Cognitive and Affective Neuroscience*, 8 (5), 595-601. [10.1093/scan/nss036](#)
- Nieuwenhuis, S., Ridderinkhof, K. R., Blom, J., Band, G. P. & Kok, A. (2001). Error-related brain potentials are differentially related to awareness of response errors: Evidence from an antisaccade task. *Psychophysiology*, 38 (5), 752-760. [10.1111/1469-8986.3850752](#)
- Nisbett, R. E. & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological review*, 84 (3), 231-259.
- Nussinson, R. & Koriat, A. (2008). Correcting experience-based judgments: The perseverance of subjective experience in the face of the correction of judgment. *Metacognition and Learning*, 3 (2), 159-174. [10.1007/s11409-008-90](#)
- Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences*, 12 (6), 237-241. [10.1016/j.tics.2008.02.014](#)
- Ortony, A., Clore, G. L. & Foss, M. A. (1987). The referential structure of the affective lexicon. *Cognitive Science*, 11 (3), 341-364. [10.1207/s15516709cog1103_4](#)
- Pacherie, E. (2008). The phenomenology of action: A conceptual framework. *Cognition*, 107 (1), 179-217. [10.1016/j.cognition.2007.09.003](#)
- Plutchik, R. & Ax, A. F. (1967). A critique of determinants of emotional state by Schachter and Singer (1962). *Psychophysiology*, 4 (1), 79-82. [10.1111/j.1469-8986.1967.tb02740.x](#)
- Prinz, J. (2004). *Gut reactions. A perceptual theory of emotions*. Oxford, UK: Oxford University Press.
- Proust, J. (2000). Awareness of agency: Three levels of analysis. In T. Metzinger (Ed.) *The neural correlates of consciousness* (pp. 307-324). Cambridge, MA: MIT Press.
- (2009). The representational basis of brute metacognition: A proposal. In R. Lurz (Ed.) *Philosophy of animal minds: New essays on animal thought and consciousness* (pp. 165-183). Cambridge, UK: Cambridge University Press.
- (2013). *The philosophy of metacognition. Mental agency and self-awareness*. Oxford, UK: Oxford University Press.
- Reber, R., Fazendeiro, T. A. & Winkielman, P. (2002). Processing fluency as the source of experiences at the fringe of consciousness. *Psyche*, 8 (10), 1-21. [10.1155/6152](#)
- Reber, R., Schwarz, N. & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and Social Psychology Review*, 8 (4), 364-382. [10.1207/s15327957pspr0804_3](#)
- Recanati, F. (2002). Does linguistic communication rest on inference? *Mind & Language*, 17 (1-2), 105-126. [10.1111/1468-0017.00191](#)
- Schwartz, B. L., Travis, D. M., Castro, A. M. & Smith, S. M. (2000). The phenomenology of real and illusory tip-of-the-tongue states. *Memory & Cognition*, 28 (1), 18-27. [10.3758/BF03211571](#)

- Schwarz, N. (2004). Metacognitive experiences in consumer judgment and decision making. *Journal of Consumer Psychology*, 14 (4), 332-348. [10.1207/s15327663jcp1404_2](https://doi.org/10.1207/s15327663jcp1404_2)
- Schwarz, N. & Clore, G. L. (2007). Feelings and phenomenal experiences. In A. W. Kruglanski & E. T. Higgins (Eds.) *Social psychology: Handbook of basic principles* (pp. 385-407). New York, NY: Guilford.
- Smith, E. R. & De Coster, J. (1999). Associative and rule based processing. In S. Chaiken & Y. Trope (Eds.) *Dual-process theories in social psychology* (pp. 323-336). New York, NY: Guilford.
- Solomon, R. C. (2007). *True to our feelings: What our emotions are really telling us*. New York, NY: Oxford University Press.
- Stepper, S. & Strack, F. (1993). Proprioceptive determinants of emotional and nonemotional feelings. *Journal of Personality and Social Psychology*, 64 (2), 211-220. [10.1037/0022-3514.64.2.211](https://doi.org/10.1037/0022-3514.64.2.211)
- Strawson, P. F. (1959). *Individuals*. London, UK: Methuen.
- Thompson, V. A., Prowse Turner, J. A. & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, 63 (3), 107-140. [10.1016/j.cogpsych.2011.06.001](https://doi.org/10.1016/j.cogpsych.2011.06.001)
- Tye, M. (2008). The experience of emotion: An intentionalist theory. *Revue Internationale de Philosophie*, 243 (1), 25-50.
- Unkelbach, C. (2007). Reversing the truth effect: Learning the interpretation of processing fluency in judgments of truth. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33 (1), 219-230.
- Unkelbach, C. & Greifeneder, R. (2013). A general model of fluency effects in judgment and decision making. In C. Unkelbach & R. Greifeneder (Eds.) *The experience of thinking: How the fluency of mental processes influences cognition and behavior* (pp. 11-32). London, UK: Psychology Press.
- Whittlesea, B. W. & Williams, L. D. (2000). The source of feelings of familiarity: The discrepancy-attribution hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26 (3), 547-547. [10.1037/0278-7393.26.3.547](https://doi.org/10.1037/0278-7393.26.3.547)
- (2001). The discrepancy-attribution hypothesis: I. The heuristic basis of feelings and familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27 (1), 3-13. [10.1037/0278-7393.27.1.3](https://doi.org/10.1037/0278-7393.27.1.3)
- Winkielman, P. & Cacioppo, J. T. (2001). Mind at ease puts a smile on the face: Psychophysiological evidence that processing facilitation elicits positive affect. *Journal of Personality and Social Psychology*, 81 (6), 989-1000. [10.1037/0022-3514.81.6.989](https://doi.org/10.1037/0022-3514.81.6.989)
- Wolpert, D. M., Ghahramani, Z. & Flanagan, J. R. (2001). Perspectives and problems in motor learning. *Trends in Cognitive Sciences*, 5 (11), 487-494. [10.1016/S1364-6613\(00\)01773-3](https://doi.org/10.1016/S1364-6613(00)01773-3)
- Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, 9 (2 pt 2), 1-27. [10.1037/h0025848](https://doi.org/10.1037/h0025848)
- (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist*, 35 (2), 151. [10.1037/0003-066X.35.2.151](https://doi.org/10.1037/0003-066X.35.2.151)

The Extension of the Indicator-Function of Feelings

A Commentary on Joëlle Proust

Iuliia Pliushch

In the following commentary I will first briefly review the target article, then voice some critical points, and last offer a positive proposal according to which tension in self-deception is a kind of a metacognitive feeling. Proust offers a novel, inspiring view that feelings possess an indexical (non-conceptual) format, are transparent (that is, they may be re-described in propositional terms, but not thereby changed), and acquire valence if the rate of change towards fulfilling the given affordance is greater or less than expected. In my critique I will first point to difficulties in disentangling feelings from emotions, then try to provide a more precise description of the formal object of feelings, along with some examples, and offer a definition of “directness” that is consistent with predictive coding—as well as argue that feelings might be influenced by concepts even if they themselves are non-conceptual. Last, I propose that tension in self-deception is a metacognitive feeling.

Keywords

Affective feelings | Appraisal | Metacognitive or noetic feelings | Predictive coding | Self-deception | Two-factor account

Commentator

[Iuliia Pliushch](#)

pliushi@students.uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Target Author

[Joëlle Proust](#)

joelle.proust@ehess.fr
Ecole Normale Supérieure
Paris, France

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 The expressive mode of feelings

First, I would like to repeat, in short, the main claims of the target paper that will serve as a basis for my subsequent comments and extensions in the following sections. Joëlle Proust’s article is concerned with the functional and informational characterisation of feelings. She argues that the concept of “feeling” consists of the following components:

1. Reactive (associated with appraisal)
2. Subjective experience

3. With distinctive embodied phenomenal quality (somatic markers have the function of expressing intensity and valence of feelings, [Proust this collection](#), p. 8)
4. Possessing a formal object (not always, e.g., feeling depressed; absence of a formal object is typical of moods, footnote 5)

The formal object of feelings is argued to be affordance-sensing, a “non-conceptual and entirely subjective appraisal of the environment by the

agent” (Proust [this collection](#), p. 7) or a “*subjective, embodied relation* to an opportunity in an input from the environment” (p. 8). Assuming the non-referential indexicality of feelings, or that feelings signal a *relational* affordance (p. 7) that depends on the representational system (p. 8), Proust argues that feelings can misrepresent only if they are re-described in propositional terms. She argues that feelings are *transparent*, because of the experienced connection between their somatic markers and affordances, as well as because of the direct influence of their valence and intensity on an agent’s motivation (p. 12). Though subjects feel directly, in order to report their feelings they have to “translate one mode of representation into another, with no guarantee that this translation will not enrich or modify FS intentional content” (p. 12). Subjects might reinterpret and mis-describe their feelings, but they cannot thereby change the nature of those feelings (feelings being *cognitively impenetrable*; p. 19).

Feelings are argued to be a plausible candidate for a natural kind on the basis of the comparison between feelings and emotions—which she considers not to constitute a natural kind (Proust [this collection](#), p. 3). Two kinds of subjective appraisal might be part of an emotion: *primary feelings* on the one hand and *appraisals cum conative dispositions* on the other. While the first kind corresponds to an earlier time in our evolutionary development, is independent of concepts, induces specific responses, and possesses distinct somatic markers, the second kind is not and might be a blend of different instances of the first kind. Apart from primary affective feelings, somatic, agentic, and metacognitive feelings are argued to form a natural kind.

The function of feelings is to non-conceptually evaluate and signal the result of a comparison process between prediction and outcome through embodied experience (Proust [this collection](#), p. 4). Due to their non-conceptual monitoring nature, feelings do not convey, but merely approximate a causal relation between internal states, external states, and actions (p. 11). There are three kinds of functional relations between feelings and actions (pp. 4–5):

1. Determination of a kind of action in response: approach vs. avoidance
2. Specific orientation in time: predictive vs. retrospective
3. Level of arousal: elevation in energy vs. soothing effect

Feelings are argued to be the result of a *comparator* or control mechanism that is *automatic* and *encapsulated*. The latter requirements are imposed in order to explain the independence of feelings of beliefs and intentions (p. 5) such that, e.g., one could still feel the adrenalin rush even though the hypothesized venomous snake turned out to be a twig.

Metacognitive feelings (M-feelings) are held to express informational, instead of environmental affordances, arise in mental acts, and trigger similar actions of approach or avoidance. M-feelings involve appraisal of the properties of the informational processes underlying contents of thought, but not those content themselves. Against Schachter & Singer’s (1962) two-factor theory of emotions (interpreted as feelings possessing intrinsic arousal but extrinsic valence), Proust argues that feelings have *intrinsic intensity and valence*. Cues on which those feelings are based can be conveyed verbally though, and thus, the heuristics (implicitly or explicitly) might change in the long run. The main claim is thus that *context-dependency is not concept-dependency* (Proust [this collection](#), p. 17). Experience of tokens of feelings differs with respect to the kind of affordance they express (several affordances might be linked to the same type of feeling) and actions they trigger.

An especially interesting claim for me is that affective feeling in general, and metacognitive feelings in particular, have a meta-monitoring function of signalling “the rate of reduction of discrepancy toward a confidence threshold” (Proust [this collection](#), p. 21). If the rate of discrepancy reduction is above expected, the valence of a feeling is experienced as more positive, and, if below expected, as more negative. “Cold” feelings without valence are those for which the expectation has been correct. This claim is interesting for two reasons. On the one hand, to the reader familiar with the self-decep-

tion literature the key-concept “confidence threshold” will stand out. It plays an important role in accounts of self-deception that regard it as a kind of hypothesis testing (one prominent proponent of this view is [Mele 2012](#)). In short, according to this type of account, gathering of evidence in favour of a certain hypothesis is pursued up to a certain point: up until the amount of evidence has reached a confidence threshold that is enough to push an acceptance or rejection of the hypothesis (for more see [Pliushch & Metzinger 2015](#)). On the other hand, “prediction error”, or difference between prediction and sensory input, is the key-term in the model of mental representation that has lately gained a large amount of acceptance—predictive coding (for a short introduction to the free-energy principle of which predictive coding is a particular implementation see [Friston 2009](#); see also [Clark, Hohwy, Seth this collection](#)). Predictive coding provides a unifying explanation for perception, cognition, and action as a result of hierarchical Bayesian inference: at different levels, predictions are compared to propagated precision-weighted prediction error that, under different conditions, leads either to changes in the model of causes of sensory input or to action directed at testing the current model ([Clark 2013](#)).

The idea that feelings signal the rate of reduction of prediction error might be worth elaborating in the predictive coding framework, particularly given the recent study by [Furl et al. \(2010\)](#) who argue that facial expressions are represented as anticipated *trajectories* of the change of those expressions: pictures of neutral and fearful faces were morphed to different degrees such that participants got to see trajectories from a neutral to a fearful face and vice versa. After seeing such a sequences of pictures, participants had to rate another picture for fearfulness. The results indicated that predictable sequences in which the degree of being morphed rose or fell monotonously, thus forming a trajectory, biased perception ([Furl et al. 2010](#), p. 696). Combining Proust’s idea with the results of Furl et al.’s study: feelings might also be represented as anticipated trajectories of change, particularly given the possibly bi-direc-

tional causal influence between feelings and facial expressions (see section 2.2).

2 Critique: Affect and implicit heuristics in feelings

2.1 Use of the term “affect”

The aim of this section is threefold: 1) show difficulties in disentangling feelings from emotions; 2) attempt to give a more precise characterisation of the formal object of feelings, along with some examples; 3) criticize the use of the term “direct” and offer another definition that is consistent with predictive coding. The first problematic point that I see is Proust’s use of the term “affective”, which is ambiguous. She employs at least two different definitions of “affective”:

1. Feelings that possess valence (p. 20). Yet *all* kinds of feelings, according to Proust, possess affect *and* valence¹ (p. 1). Given her distinction between “hot” (emotional) feelings and those that have valence² (p. 21), emotional feelings might differ from mere feelings with valence due to the differently-experienced valence, maybe if emotional valence were a richer experience. Thus, the question is about the minimal requirements on valence and intensity in feelings.
2. Feelings that express emotions.
3. *Difference between feelings and emotions*: if agentive and metacognitive feelings can be affective, then the categorization of feelings into bodily, agentive, metacognitive, *and* affective (p. 5) might be better restricted to the first three, with the fourth being a dimension along which they vary. If affective in this categorization means emotional (p. 2), then there is an ambiguity of terms—affect-

1 The following quotations might help to elucidate the matter: “[f]eelings typically express affect and valence in sensation (25-26), all the feelings vary in affect in roughly the same way, because they all include valence in their informational structure” (p. 20).

2 In Proust’s words, the difference between “hot” feelings and feelings with valence, on the example of M-feelings, is that “although all M-feelings do not often have a definite ‘hot’ quality comparable with fear and love, they always have a valence, according to whether they predict the agent’s progress toward or away from her cognitive goal” (p. 21).

ive = having valence and affective = being part of an emotional experience—because the latter seems to be more complex.

4. *Difference between formal objects of feelings and emotions*: if “feelings are affective ingredients in emotional awareness” (p. 3), then there is a circularity in understanding affectivity here: feelings are affective in virtue of being part of an emotion, while at the same time they themselves are the affective component in the emotion of which they are part. The first part of this claim can be followed from that defended by Proust, namely that feelings that do not express emotions are not affective (p. 2). The second part of the claim follows from Proust’s claim that feelings are affective ingredients of emotions (p. 3). As elaborated in the previous section, emotions are said by Proust to contain one of two kinds of subjective appraisals: feelings or appraisals cum conative dispositions. Further, if feelings are components of emotions, but both can have a formal object, then those objects might diverge. The consequence is that an emotion and a feeling that is part of it might be directed at different objects. Thus, Proust on the one hand distinguishes feelings from emotions and yet on the other hand claims that not only emotional feelings, but also agentive and metacognitive feelings might be “feeling toward” experiences (p. 3, pp. 20–21). The latter claim that both feelings and emotions are directed at intentional objects has been used as an argument to identify both (see [de Sousa 2014](#) section 2 for a discussion of this question). Given Proust’s claim that there are somatic, affective, agentive, and metacognitive feelings, and given the claim that at least in metacognitive feelings the formal object is not the cognitive disposition itself but the rate of change of its execution above or below discrepancy, an interesting question focuses on the formal object of emotional feelings.³ For example, can it be

³ To be more precise, the question is about the functional description of the formal object of feelings. Proust ([this collection](#)) says that “[f]eelings express [...] affordance as their focus (for formal object), along with its graded valence, ranging from very unpleasant to very pleasant, and with its intensity gradient, which ranges from small to large” (p. 8). Affordance is defined as “perceived utility”, and can be

that while the formal object of the emotion of fear is some dangerous object, the object of a feeling is a rate of change in the assessment of the situation before and after the change of the formal object of an emotion? This might explain why, e.g., the first bite of a bar of chocolate makes one happier than the following bites.

5. *Bodily phenomenology of feelings as their formal object*: Proust argues that while somatic feelings are about bodily sensations (or, more consistently, about the rate of their change), in affective (emotional) and possibly metacognitive feelings “the bodily phenomenology tends to recede to the fringe of consciousness” ([this collection](#), p. 2). The example that Proust gives with respect to metacognitive feelings is that feelings of remembering are correlated with but not about facial muscle activity (p. 3). Proust acknowledges that there might be mixed cases (experience of bodily feeling + intentional content, pp. 2–3), but I want to argue that in some emotional feelings bodily phenomenology is, to borrow a metaphor, in the *foreground*. There might be emotional feelings whose objects are bodily sensations, e.g., the anxiety that arises during a panic attack: when I concentrate on my accelerated heartbeat, then if I come to associate the heartbeat with some threatening aspects of a situation, such an experience might lead to anxiety, and thus the initial anxiety leads to even more anxiety, leading to a vicious cycle of panic (for a discussion of heartbeat perception in panic disorder see [Ehlers & Breuer 1996](#)). This might be a case of an emotion whose formal object is the rate of change of bodily sensations, or maybe a meta-feeling (for a discussion of meta-emotions see [Mendonça 2013](#)).

In the given panic example it might have seemed as if I had embraced the analogy between feelings and perception that Proust

positive or negative (*ibid.*, p. 7). Positivity and negativity are dimensions along which valence changes, and valence has been characterised as the rate of change of discrepancy towards the (cognitive) goal. For more on why the latter characterisation is interesting see section 3.

denies, so I will explain why it may be more beneficial to use the term “direct” in another sense to that used by Proust. Proust makes a sharp distinction between feeling and perceiving: “[w]hile percepts allow recognition and identification of external objects and properties, feelings express specific affordances in a perceived, imagined, or remembered situation” (this collection, p. 10). Non-conceptual parts of perceptions are said to relate to “objective, external contrastive cues” (Proust this collection, p. 10), while in feelings they relate to evaluative states. Perception is said to involve “direct sensory access to the world” (p. 10), while the access of feelings to the world and the body is claimed to be indirect. Proust’s evidence for a disanalogy between feeling and perception is based on the neuroscientific research of Barrett & Bar, who say that absence of “internal affective context” impairs the categorization of objects (2009, p. 12).⁴ Their evidence for this hypothesis is based on reviewing the anatomic connections involved in affective processing and that of object perception. One critique of this might be that the time of activation of certain regions responsible for emotional processing and perception might justify the claim that emotional processing comes before perception, but not how direct such processing is. Moreover, in light of predictive coding, perception, emotion, and cognition might all be indirect (Hohwy 2014; for more technical elaboration Friston et al. 2014). In other words, predictive coding provides the term “direct” with a meaning other than that used by Proust. In predictive coding directness is an absence of the evidentiary boundary, where the evidentiary boundary is the inferential isolation between the model of the world and the hidden causes of sensory input (Hohwy 2014). This means that causes beyond the boundary have to be inferred on the basis of independent evidence (ibid., p. 6), or, in Hohwy’s words, “[t]he brain doing the inference is secluded at least in the sense that certain

kinds of doubt about the occurrence of the evidence are unanswerable without further, independent evidence” (p. 7). Relating this observation to Proust, on the premise of accepting predictive coding, there might not be a sharp distinction between feeling and perceiving such as Proust postulates, or at least not in the form presented in the target article. If interoception as perception involves inferences about circumstances *beyond* the (same) evidentiary boundary, as suggested by Hohwy (2014), then feeling and perceiving would both be indirect (to the same degree).⁵ If interoception does not go beyond the evidentiary boundary, feelings might be direct, even if perception is not.

2.2 Concept-based feelings?

In this part of the review I will point out the dangers of interpreting the relation between feelings and concepts too simplistically and argue that it is possible that at least some kinds of feelings are influenced by concepts, even if they themselves are non-conceptual. Proust argues that for metacognitive feelings to arise an important affordance, as well as an implicit heuristic, has to be present (this collection, p. 18). This heuristic is based on cues about the dynamics of information processing, but not its contents (p. 15). The dichotomies that Proust uses in the description—implicit–explicit, unconscious–conscious, evolutionarily-old–evolutionarily-new, associative–rule-based (pp. 3–4, p. 17)—have often been mapped onto two different kinds of processes in dual processing theory (e.g., Frankish & Evans 2009). Dual processing theory states that there are two kinds of processing that possess the dichotomous characteristics mentioned above. A minimal description provided by Evans (2009) for type 1 is “fast, automatic, high processing capacity, low effort”, and for type 2 “slow, controlled, limited capacity, high effort” (p. 33). Along these lines, “implicit”, “unconscious”, “evolutionarily old”, “associative” have been also used as descriptors for type 1 and “explicit”, “conscious”, “evolutionar-

⁴ Barrett & Bar (2009) define affect as an influence on bodily states that is either unconscious or, if conscious, experienced as pleasurable or unpleasurable to varying degrees (pp. 1327–1328). Barrett & Bar’s (2009) basic claim is that the orbitofrontal cortex (OFC) integrates into a unified multimodal representation sensory information from both world and body in a dynamic way.

⁵ One could also ask whether the *same* evidentiary boundaries would be involved in feeling and perceiving, since there could be many of them (Friston 2013).

ily new”, “rule-based” as descriptors for type 2. A belief bias (accepting more believable than unbelievable conclusions) might serve as an example for type 1 (*ibid.*, p. 41), and the conscious correction thereof for type 2. The worry I have is adding to those dichotomies another one: non-conceptual (meaning in this case non-propositional; Proust *this collection*, p. 7)–conceptual (propositional, belief-like). Proust holds that “cues (associative heuristics) dictate how an affordance is detected, assessed and exploited in a context, but these cues are not consciously available, and hence do not depend on a naïve theory of the task” (p. 17). This inference is not valid in the given form. I agree with Proust that “[a] cue-based, non-analytic heuristic is not inferential in the interpretive, first-person sense” (p. 17), but I hold that there is at least one step to consider in between non-conceptual⁶ affordances and consciously evaluated affordances. And this is automatic *concept-based* activation (the existence of automatic appraisal is acknowledged by Proust; footnote 7).

Evans (2009) distinguishes between different kinds of dual processing theories, among which are the sequential (first automatic processing, then controlled) and the parallel theory. Proust seems to embrace a sequential kind of dual processing theory, given the functional role she ascribes to metacognitive feelings (evaluation of mental actions before and after their execution; Proust 2013). Yet how far implicit heuristics are independent of concepts is in question. Proust (*this collection*) denies that “a concept-based interpretation will affect the experienced feeling itself” (p. 17). As mentioned in section 1, she also denies that feelings have a conceptual format. Thus, she seems to deny both that concepts play a causal role in the emergence of feelings and that feelings themselves possess a conceptual format. I will briefly demonstrate that the term “implicit heuristic” does not preclude automatic concept activation, if it implies the activation of knowledge or goal representations. Thompson (2009) argues that

heuristic processes are contaminated by background knowledge, as well as by beliefs and expectations (p. 172, p. 174). Frankish (2009) notes that “the concepts of belief and desire correspond to the psychologist’s concepts of knowledge (or memory) and goal structure” (p. 91). Hence, activation of knowledge that may provide the context for feelings could also be conceptual. Goal representations might also be activated in the course of context creation, provided that unconscious goal pursuit is flexible and context-sensitive (Aarts & Custers 2012). Further, unconsciously activated goals not only depend on context, but also *create* context by influencing the accessibility of knowledge, evaluations, and emotions (Fishbach & Ferguson 2007, p. 496). It follows that if goal representations are activated, then they might lead to the activation of conceptual knowledge. Another interesting point is that if there is a continuous interplay between goal representations and affordances (opportunities in the environment; Huang & Bargh 2014, p. 125) and if goal representations can change the experience of the world (*ibid.*, p. 124), then goal representations might change sensing of affordances and, hence, the feelings associated with it. Further, there has been a proposal to distinguish between associative and rule-based processes by the kind of architecture they operate upon: namely connectionist vs. classical computational (for a short discussion see Samuels 2009, pp. 141–142). Thus, implicit heuristics might be understood as certain connected representations in a network being activated by some cues, where the question is about the representational format of such knowledge, or a more precise description of the relational nature of the feeling affordance. Last, a general note about the similarity between feelings and other kinds of representations: if Bliss-Moreau & Williams (2014) are correct in defending the claim that all kinds of representations possess an affective component (valence + arousal in their definition), then affect is something that expressive and conceptual representations share.

Of course, Proust’s claim that in the case of feelings those cues relate to the dynamics, but not to the contents of processes, indicates a

⁶ Among those who agree with Proust that the content of epistemic feelings is non-conceptual and non-metarepresentational are, for example, Michaelian & Arango-Muñoz (2014). But the *content* of a metacognitive feeling being non-conceptual does not preclude that concepts play a causal role in its emergence.

more specific understanding of the kind of implicit heuristic in question. My point, though, is that if humans can “enrich their noetic feelings through concepts, and thereby revise their reliance on fluency where it is not justified” (Proust 2013, p. 144), then in humans implicit heuristics may also be influenced by concepts (in an automatic way) and in such a way influence feelings. Needless to say, the independent existence of such a schema (be it cognitive or emotional) is hard to prove (Eysenck & Keane 2010, p. 597). According to Koriat & Levy-Sadot (1999), as cited by Proust (this collection, p. 15), metacognitive feelings arise as a result of nonanalytic inferential processes (described as the implicit or unconscious application of heuristics), in distinction to the direct memory trace hypothesis, according to which feelings have direct access to memory traces (Koriat & Levy-Sadot 1999, p. 487). Koriat & Levy-Sadot (1999) argue that the presence of dissociations between knowing and the feeling of knowing speaks against the second hypothesis. Even if heuristics in feelings are non-conceptual, the fact that through feelings emotion gets its valence necessitates that we consider how concepts and memory traces influence feelings, given that they play a role in emotions. Lane et al. (forthcoming), for example, argue that psychotherapeutic change is made possible by updating prior emotional experiences, for which memory traces of those experiences have to be reactivated and reconsolidated. Thus, even if feelings are non-propositional (Proust this collection, p. 20), activation of concepts and their expression in propositional terms are to be distinguished. The point is not that metacognitive feelings themselves cannot have indexical formats,⁷ or that an agent could not possess expressive and conceptual representations at the

same time, but that in humans the generation of (at least) metacognitive and emotional feelings might be preceded by an automatic concept activation that influences them. If this is the case, then one could ask again whether feelings are transparent (see section 1).

Further, instead of describing cognitive processes as serial, their *dynamic* (continuous) nature might be more worthy of emphasis. In the target article, Proust mentions that “[i]ncreased activity in the smile muscle, the zygomaticus major, produces feelings with a positive valence” (this collection, p. 15). This suggests that facial expression influences emotions. She also argues for the transparency (impenetrable nature) of feelings and the against two-factor theory, thus against the possibility that appraisal influences the valence of feelings (see section 1). I want to offer for clarification purposes a short review of the recent literature on which factors are supposed to influence feelings and factors feelings influence themselves. Rogers et al. (2014) emphasize the dynamic nature of emotions insofar as they depend on the social appraisal of a situation. Brosch (2013) also emphasizes the dynamic nature of appraisal that plays a causal role in eliciting emotions. The definition of appraisal that Brosch (2013) accepts also encompasses low-level appraisal based on learned schemata (p. 370). Brosch (2013) argues that first an initial low-level appraisal affects the physiology (1), action tendency (2), expression (3), and feeling (4) of an emotional experience, and then those changes in turn affect an on-going (low- and high-level) appraisal, establishing an appraisal loop. Here, the direction of influence is still in question, e.g., whether feelings influence expressions or the other way around. Laird & Lacasse (2014) defend the James-Lange theory of emotion, namely that facial expressions (e.g., BOTOX patients being less responsive to mild positive emotional stimuli; for the reference see *ibid.*, p. 29), expressive behaviour (e.g., romantic attraction as a result of shared, mutual gaze; *ibid.*, p. 29), and visceral responses that are interpreted according to situational cues (e.g., misattribution of emotion) are *causes* of emotions (for a critique of their evidence see Reisenzein &

⁷ A better understanding of the indexical mode of feelings might be provided by the following quotation: “Feelings can be seen as pre-specified states of a comparator, which predict ultimate success or failure in the actions that they monitor. Given that the information they carry is immediately used in controlling and monitoring current effort, it is misleading to present them as ‘reporting’ the epistemic properties of a mental state or referring to it (even *de re*). They are, rather, signals in a control mechanism, which work somewhat as traffic lights do: allowing traffic, stopping it, rechanneling it; no report or reference need be involved” (Proust 2013, p. 76). In another place Proust (2013) notes that feelings “do not properly ‘refer’, because they do not engage propositional thinking” (p. 77).

Stephan 2014). As such, they may influence emotional feelings too, which Proust acknowledges by pointing out the causal connection between measures in facial muscles and affective character of feelings (this collection, p. 25). Yet the direction of influence may also go the other way around (from feelings to facial expressions). Thus, the nature of feelings may also be dynamic, as are the nature of the underlying cognitive processes. Interestingly, Thagard & Schröder (2014) argue for a neurocomputational theory of emotions as semantic pointers (term introduced by Chris Eliasmith). They argue that physiological, appraisal, social, and psychological components of emotions can be integrated into one unified account: emotion tokens can possess both shallow and deep meanings. The compressed (shallow) form of emotions is reportable, while at the same time pointing to the uncompressed deep form that binds together situational, physiological, and appraisal components.

In the preceding paragraph I considered literature supporting the claim that feelings are embedded in *continuous* cognitive processes. The purpose of this was to show that how appraisal might influence feelings in some form is complex and might even be circular. In this paragraph I offer some additional evidence against a discontinuous interpretation of the connection between feelings and propositional descriptions thereof. The existence of *affective blindsight* (ability to discern emotional stimuli despite inability to consciously perceive them; Eysenck & Keane 2010, p. 581) would stand in line with the assumption that emotional and cognitive processing is based on different kinds of information. This is because affective blindsight demonstrates the dissociation between two different kinds of processing and, thus, a dissociation between the information needed for the one kind and for the other. Further, Scott et al.'s (2014) experiment demonstrating *blind insight* (accurate metacognitive accuracy in the absence of discriminative accuracy) on the one hand supports Proust's hypothesis that metacognition and first-order cognition are not based on the same kind of information, yet on the other it speaks against a serial interpretation

according to which feelings arise out of automatic processes and are then re-described in propositional terms and used in first-person inferential reasoning. Liu & Wang (2014), for example, argue that motivational intensity influences the effect of positive affect on cognitive control: low-approach motivated positive affect enhances cognitive flexibility and distractibility, while high-approach motivated positive affect (associated with goal pursuit) enhances cognitive stability. Thus, the role of feelings might be broader than just the indicators that may or may not be used in conscious reasoning.

3 Proposals: Tension in self-deception is a kind of metacognitive feeling

Proust (this collection, as well as 2013) argues that mental actions are preceded and followed by metacognitive feelings indicating the appropriateness of the cognitive process in question. I want to argue that tension in self-deception fits the characterisation of a metacognitive feeling. Tension is described as a feeling of uneasiness and distress, and as such I think that it is precisely this tension that is said to indicate to the self-deceiver that her belief-forming process is faulty.

Self-deception (SD) is a motivated (1) kind of typically subpersonal hypothesis-testing (2) that results in an evidence-incompatible mental representation of reality (3) which fulfils a belief-like role (4) (Pliushch & Metzinger 2015). Self-deception is usually discussed in the context of biased belief-forming processes and it is argued that phenomenological tension arises as a result of the execution of such processes (e.g., Lynch 2012). Thus, the same function has been ascribed to tension in self-deception as the one ascribed by Proust to metacognitive feelings, namely a comparison of the cognitive process to certain criteria. In self-deception, rationality criteria are typically emphasised.

I want to argue that metacognitive feelings apply to self-deception, insofar as they might also monitor *unconscious* cognitive processes and arise not only before or after a cognitive process, but also *during* it. In case of self-deception these cognitive processes are belief-forming

processes. Proust (this collection, 2013) considers conscious mental actions: her argument is that unconscious comparison processes that give rise to metacognitive feelings precede and follow conscious mental actions. She argues that the “attentional-supervisory system” emerges from “distributed metacognitive abilities” (Proust 2013, p. 263). Ignorance of epistemic norms such as relevance, coherence, fluency, and informativeness lead to (pathological) errors in belief acquisition (Proust 2013, pp. 260–261). My argument in favour of the extension of metacognitive feelings to monitor unconscious cognitive processes is of a phenomenological nature. I agree with Proust (this collection) that the term “inference” has been used loosely in the literature and does not always indicate a first-person inference (p. 21). Yet the more basic problem might be that there is no sequential first-person inference as such in the first place. If the shift between mind wandering (task-unrelated cognitive activity) and task-directed cognitive activity goes unnoticed (Metzinger 2013), then there might be other shifts that we do not notice, e.g., the shift from unconscious to conscious cognitive processes, or some changes in the given process. Thus, the phenomenology of a cognitive process might be more complicated than a unified sequence with a starting point and an end. Further, given, for example, mood-state dependent cognition (Eysenck & Keane 2010, pp. 584), I doubt the plausibility of the assumption that only in breaks between conscious cognitive processes do subjects experience affective feelings.

In the previous paragraphs I argued that the functional role of metacognitive feelings fits that of tension in self-deception, and that metacognitive feelings arise not only before and after mental actions, but also before, after, and during unconscious (possibly self-deceptive) cognitive processes. In this paragraph I want to link Proust’s idea that feelings possess valence only if the rate of change of progress is unexpected to predictive coding, in order to provide a functional description of metacognitive feelings. Proust (this collection) argues that the affective quality of feelings arises only if the cognitive process violates expectations: if it progresses

quicker towards the goal, positive feelings arise, if slower, negative feelings arise⁸ (p. 21). Given that the terms “expectation” and “prediction error” have gained popularity in virtue of being key terms in predictive coding, which is a modelling strategy explaining perception, cognition, and action (Clark 2013), I will shortly discuss Proust’s claim about affect in metacognitive feelings in the context of predictive coding. According to predictive coding, prediction errors (deviation between expectation and outcome) are precision-weighted. Precision is the property of prediction errors (errors between the top-down prediction and the bottom-up signal one receives) that can be described as the weight of a prediction error that plays the role of selection: the more precise the prediction error, the more it will change the hypothesis about causes of input. Switching between perception and action depends on the precision of prediction errors: precise prediction errors change hypotheses, while imprecise ones lead to action (Brown et al. 2013). Precision⁹ is also argued to play a dual biasing role: biasing perception toward goal states and enhancing confidence in action choices (Friston et al. 2013). Low precision of prediction errors has been argued to cause anxiety (Mathys et al. 2011, p. 17).¹⁰ I argue that Proust’s proposal that violations of expectations of “a given rate of reduction of the discrepancies toward her [agent’s] cognitive goal” (this collection, p. 26) produce affective feelings might be described in predictive coding terms as violations of *transition probabilities* of reaching the goal state:¹¹ if a state conducive to the goal state or a goal state itself has been reached, despite a low probability of changing

8 Note the analogy to the “dark room problem” in predictive coding: if an agent wants to minimize surprise or prediction error, then she should stay in a dark room, given that there will be no surprise in it (e.g., Clark 2013). If there were no prediction error, this would cause uncertainty (e.g., Friston et al. 2012). Proust’s argument is similar: if there were no violations of expectations, then metacognitive feelings would not have any valence, because they only have valence if the rate of change is quicker or slower than expected.

9 Attention is precision optimization according to predictive coding (Hohwy 2013).

10 Mathys et al. (2011) are also interesting for the given topic insofar as Proust argues that the heuristics upon which metacognitive feelings are based might be changed via associative learning; Mathys et al. (2011) provide a predictive coding model of reinforcement learning.

11 For a predictive-coding model of a goal-directed action see Friston et al. (2013).

into that state from the current state, then positive affective feelings might arise.¹²

The first step in the categorisation of tension as a metacognitive feeling has been an extension of the application of metacognitive feelings to unconscious belief-forming processes. The second is to clarify the representational content of tension. To do the latter, it might be beneficial to consider which other kinds of metacognitive feelings arise out of belief-forming processes. Those are intuitivity, counter-intuitivity, and anxiety, if one classifies them according to the phenomenology and not according to the norm that they control. Intuitivity indicates the appropriateness of a given belief-forming process.¹³ The reason for the ascription of the given functional role to intuitivity is that intuitivity signals 1) a good fit with respect to the network of our explicit background beliefs and 2) a good fit with respect to our conscious and unconscious model of reality (Metzinger & Windt 2014). An appropriate belief-forming process provides a good fit with respect to 1) and highly likely also with respect to 2). I further argue that counter-intuitivity represents that a certain cognitive process violates the chosen criterion of appropriateness, but is neutral with respect to the system's goal representations, while tension or anxiety represents that the cognitive process violates at least some important goal representations. The reason for this distinction is to account for the effect of motivation on belief-forming processes.

Thus, if feelings accompany our belief-forming processes, then readers might have experienced some while reading this commentary: hence the title. To conclude, I think that Proust has offered interesting ideas on the nature of feelings that will greatly contribute to the clari-

fication of the matter: the indexical (affordance-sensing and non-conceptual) format of feelings, their transparency, the taxonomy of feelings into sensory, emotional, agentive, and epistemic, the predictive and retrospective function of feelings signalling the appropriateness of the cognitive process they monitor, and the degree of change of expectation as the origin of valence of feelings. In this review I have tried to extend Proust's account. To do this, I attempted to provide some conceptual clarifications on the distinction between feelings and emotions, the formal object of feelings, and the conceptual influences to which they might be subject. Last, I argued that tension in self-deception is a kind of metacognitive feeling.

Acknowledgements

I am very grateful to the Barbara-Wengeler foundation for generously supporting my PhD project and all the reviewers of this commentary for their helpful improvements.

¹² Emotional valence has been also argued to be modelled as the rate of change of free energy: Instead of estimating volatility or “slow and continuous changes in states of the world” the rate of change of free energy is argued to take that role of estimating (known) uncertainty (Joffily & Coricelli 2013, p. 1). Here Joffily & Coricelli (2013) accept Yu & Dayan's (2005) distinction between expected and unexpected uncertainty: Expected uncertainty is the one about known unreliability of predicting relationships *within* a context and unexpected uncertainty is the one about the appropriateness of the context itself such that when unexpected uncertainty is high, it is a signal that a *context switch* should be made.

¹³ For an elaboration on the phenomenal signature of knowing in intuitions of certainty, see Metzinger & Windt (2014).

References

- Aarts, H. & Custers, R. (2012). Unconscious goal pursuit: Nonconscious goal regulation and motivation. In R. M. Ryan (Ed.) *The Oxford handbook of human motivation* (pp. 232-247). Oxford, UK: Oxford University Press.
- Barrett, L. F. & Bar, M. (2009). See it with feeling: Affective predictions during object perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364, 1325-1334.
[10.1098/rstb.2008.0312](https://doi.org/10.1098/rstb.2008.0312)
- Bliss-Moreau, E. & Williams, L. A. (2014). Tag, you're it: Affect tagging promotes goal formation and selection. *Behavioral and Brain Sciences*, 37 (2), 138-139.
[10.1017/S0140525X13001969](https://doi.org/10.1017/S0140525X13001969)
- Brosch, T. (2013). Comment: On the role of appraisal processes in the construction of emotion. *Emotion Review*, 5 (4), 369-373. [10.1177/1754073913489752](https://doi.org/10.1177/1754073913489752)
- Brown, H., Adams, R. A., Parees, I., Edwards, M. & Friston, K. (2013). Active inference, sensory attenuation and illusions. *Cognitive Processing*, 14 (4), 411-427.
[10.1007/s10339-013-0571-3](https://doi.org/10.1007/s10339-013-0571-3)
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36 (3), 181-204.
[10.1017/S0140525X12000477](https://doi.org/10.1017/S0140525X12000477)
- Clark A. (2015). Embodied prediction. In T. Metzinger and J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- de Sousa, R. (2014). Emotion. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy (Spring 2014 Edition)*.
<http://plato.stanford.edu/archives/spr2014/entries/emotion/>
- Ehlers, A. & Breuer, P. (1996). How good are patients with panic disorder at perceiving their heartbeats? *Biological Psychology*, 42 (1-2), 165-182.
[10.1016/0301-0511\(95\)05153-8](https://doi.org/10.1016/0301-0511(95)05153-8)
- Evans, J. St. B. T. (2009). How many dual-process theories do we need? One, two, or many? In K. Frankish and J. St. B. T. Evans (Eds.) *In two minds: Dual processes and beyond* (pp. 33-54). Oxford, UK: Oxford University Press.
- Eysenck, M. & Keane, M. T. (2010). *Cognitive psychology: A student's handbook*. Hove, UK: Psychology Press.
- Fishbach, A. & Ferguson, M. (2007). The goal construct in social psychology. In A. W. Kruglanski and E. T. Higgins (Eds.) *Social psychology: Handbook of basic principles* (pp. 490-515). New York, NY: Guilford Press.
- Frankish, K. (2009). Systems and levels: Dual-system theories and the personal-subpersonal distinction. In K. Frankish and J. St. B. T. Evans (Eds.) *In two minds: Dual processes and beyond* (pp. 89-107). Oxford, UK: Oxford University Press.
- Frankish, K. & Evans, J. St. B. T. (2009). The duality of mind: An historical perspective. In K. Frankish and J. St. B. T. Evans (Eds.) *In two minds: Dual processes and beyond* (pp. 1-29). Oxford, UK: Oxford University Press.
- Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, 13 (7), 293-301. [10.1016/j.tics.2009.04.005](https://doi.org/10.1016/j.tics.2009.04.005)
- (2013). Life as we know it. *Journal of The Royal Society Interface*, 10 (86), 20130475.
[10.1098/rsif.2013.0475](https://doi.org/10.1098/rsif.2013.0475)
- Friston, K., Adams, R. A., Perrinet, L. & Breakspear, M. (2012). Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology*, 3 (151).
[10.3389/fpsyg.2012.00151](https://doi.org/10.3389/fpsyg.2012.00151)
- Friston, K., Schwartenbeck, P., Fitz-Gerald, T., Moutoussis, M., Behrens, T. & Dolan, R. J. (2013). The anatomy of choice: Active inference and agency. *Frontiers in Human Neuroscience*, 7 (598).
[10.3389/fnhum.2013.00598](https://doi.org/10.3389/fnhum.2013.00598)
- Friston, K., Sengupta, B. & Auletta, G. (2014). Cognitive dynamics: From attractors to active inference. *Proceedings of the IEEE*, 102 (4), 427-445.
[10.1109/JPROC.2014.2306251](https://doi.org/10.1109/JPROC.2014.2306251)
- Furl, N., van Rijsbergen, N.J., Kiebel, S. J., Friston, K. J., Treves, A. & Dolan, R. J. (2010). Modulation of perception and brain activity by predictable trajectories of facial expressions. *Cerebral Cortex*, 20 (3), 694-703.
[10.1093/cercor/bhp140](https://doi.org/10.1093/cercor/bhp140)
- Hohwy, J. (2013). *The predictive mind*. Oxford, UK: Oxford University Press.
- (2014). The self-evidencing brain. *Noûs*.
[10.1111/nous.12062](https://doi.org/10.1111/nous.12062)
- Hohwy, J. (2015). The neural organ explains the mind. In T. Metzinger and J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Huang, J. Y. & Bargh, J. A. (2014). The selfish goal: Autonomously operating motivational structures as the proximate cause of human judgment and behavior. *Behavioral and Brain Sciences*, 37 (2), 121-135.
[10.1017/S0140525X13000290](https://doi.org/10.1017/S0140525X13000290)
- Joffily, M. & Coricelli, G. (2013). Emotional valence and the free-energy principle. *PLoS Computational Biology*, 9 (6), e1003094. [10.1371/journal.pcbi.1003094](https://doi.org/10.1371/journal.pcbi.1003094)

- Koriat, A. & Levy-Sadot, R. (1999). Processes underlying metacognitive judgments: Information-based and experience-based monitoring of one's own knowledge. In S. Chaiken and Y. Trope (Eds.) *Dual-process theories in social psychology* (pp. 483-502). London, UK: The Guilford Press.
- Laird, J. D. & Lacasse, K. (2014). Bodily influences on imotional feelings: Accumulating evidence and extensions of William James's theory of emotion. *Emotion Review*, 6 (1), 27-34. [10.1177/1754073913494899](https://doi.org/10.1177/1754073913494899)
- Lane, R. D., Ryan, L., Nadel, L. & Greenberg, L. (forthcoming). Memory reconsolidation, emotional arousal and the process of change in psychotherapy: New insights from brain science. *Behavioral and Brain Sciences*. [10.1017/S0140525X14000041](https://doi.org/10.1017/S0140525X14000041)
- Liu, Y. & Wang, Z. (2014). Positive affect and cognitive control: Approach-motivation intensity influences the balance between cognitive flexibility and stability. *Psychological Science*, 25 (5), 1116-1123. [10.1177/0956797614525213](https://doi.org/10.1177/0956797614525213)
- Lynch, K. (2012). On the "tension" inherent in self-deception. *Philosophical Psychology*, 25 (3), 433-450. [10.1080/09515089.2011.622364](https://doi.org/10.1080/09515089.2011.622364)
- Mathys, C., Daunizeau, J., Friston, K. J. & Stephan, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, 5 (39). [10.3389/fnhum.2011.00039](https://doi.org/10.3389/fnhum.2011.00039)
- Mele, A. R. (2012). When are we self-deceived? *Human-a.Mente - Journal of Philosophical Studies*, 20, 1-15.
- Mendonça, D. (2013). Emotions about emotions. *Emotion Review*, 5 (4), 390-396. [10.1177/1754073913484373](https://doi.org/10.1177/1754073913484373)
- Metzinger, T. (2013). The myth of cognitive agency: Subpersonal thinking as a cyclically recurring loss of mental autonomy. *Frontiers in Psychology*, 4 (931). [10.3389/fpsyg.2013.00931](https://doi.org/10.3389/fpsyg.2013.00931)
- Metzinger, T. & Windt, J. (2014). Die Phänomenale Signatur des Wissens: Experimentelle Philosophie des Geistes mit oder ohne Intuitionen? In T. Grundmann, J. Horvath and J. Kipper (Eds.) *Die experimentelle Philosophie in der Diskussion* (pp. 279-321). Berlin, GER: Suhrkamp.
- Michaelian, K. & Arango-Muñoz, S. (2014). Epistemic feelings, epistemic emotions: Review and introduction to the focus section. *Philosophical Inquiries*, 2 (1), 97-122.
- Pliushch, I. & Metzinger, T. (2015). Self-deception and the dolphin model of cognition. In R. Gennaro (Ed.) *Disturbed consciousness*. Cambridge, MA: MIT Press.
- Proust, J. (2013). *Philosophy of metacognition: Mental agency and self-awareness*. Oxford, UK: Oxford University Press.
- Proust J. (2015). The representational structure of feelings. In T. Metzinger and J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Reisenzein, R. & Stephan, A. (2014). More on James and the physical basis of emotion. *Emotion Review*, 6 (1), 35-46. [10.1177/1754073913501395](https://doi.org/10.1177/1754073913501395)
- Rogers, K. B., Schröder, T. & Scheve, C. v. (2014). Dissecting the sociality of emotion: A multilevel approach. *Emotion Review*, 6 (2), 124-133. [10.1177/1754073913503383](https://doi.org/10.1177/1754073913503383)
- Samuels, R. (2009). The magical number two, plus or minus: Dual-process theory as a theory of cognitive kinds. In K. Frankish and J. St. B. T. Evans (Eds.) *In two minds* (pp. 129-146). Oxford, UK: Oxford University Press.
- Schachter, S. & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, 69 (5), 379-399. [10.1037/h0046234](https://doi.org/10.1037/h0046234)
- Scott, R., Dienes, Z., Barrett, A. B., Bor, D. & Seth, A. K. (2014). Blind insight: Metacognitive discrimination despite chance task performance. *Psychological Science*, 25 (12), 2199-2208. [10.1177/0956797614553944](https://doi.org/10.1177/0956797614553944)
- Seth A. (2015). The cybernetic bayesian brain. In T. Metzinger and J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Thagard, P. & Schröder, T. (2014). Emotions as semantic pointers: Constructive neural mechanisms. In L. F. Barrett and J. A. Russell (Eds.) *The psychological construction of emotion* (pp. 144-167). New York, NY: Guilford.
- Thompson, V. A. (2009). Dual-process theories: A metacognitive perspective. In K. Frankish and J. St. B. T. Evans (Eds.) *In two minds: Dual processes and beyond* (pp. 171-195). Oxford, UK: Oxford University Press.
- Yu, A. J. & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, 46 (4), 681-692. [10.1016/j.neuron.2005.04.026](https://doi.org/10.1016/j.neuron.2005.04.026)

Feelings as Evaluative Indicators

A Reply to Iuliia Pliushch

Joëlle Proust

These responses aim at clarifying various aspects and implications of my proposal that feelings are affordance sensings. Affective quality, in the present proposal, extends beyond the domain of primary and secondary emotions to all feelings, because it results from specific features in the dynamics of valence. Feelings do not convey an explicit causal information about the world. Causal relations are, rather, implicitly represented in a felt affordance through the dynamic relations between the associated, embodied cues for location, valence and intensity and type of the affordance. Affordances are neither perceived nor inferred; they are “sensed”, which is an ability distinct from belief, whose informational input is derived from features of a perceived or interpreted situation or cognitive task. The input for an affordance sensing can well be conceptual; it is claimed, however, that even when a task is represented through concepts, the affordance-sensings elicited during the task are nonconceptual and evaluative. The relevant properties in affordance-sensings being dynamic, an interpretation of the view under discussion as being serial is resisted. Finally, Pliushch’s proposal for extending this theory to an interpretation of the feelings involved in self-deception is discussed.

Keywords

Affective feelings | Causal information | Metacognition | Noetic feelings | Self-deception | Serial vs. dynamic processes | Valence

Author

Joëlle Proust

joelle.proust@ehess.fr

Ecole Normale Supérieure
Paris, France

Commentator

Iuliia Pliushch

pliushi@students.uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Use of the term “affect”

One of the aims of this article is to try to define feelings according to their functional characteristics, when seen as all-purpose comparators. Iuliia Pliushch claims that my use of “affective feelings” is ambiguous, because they seem to be *defined* either as “feelings that possess valence”, or as “feelings that express emotions”. I am happy to accept the blame for not rephrasing in my own terms the subcategory of “affective feelings” discussed in emotion theory.

A similar discrepancy, however, may seem to be present between two passages of my chapter where I do express my own view:

As will be seen below, some feelings, however, do not express emotions, i.e., are not affective. ([Proust this collection](#), p. 2)

All the feelings vary in affect in roughly the same way, because they all include valence in their informational structure. ([Proust this collection](#) p. 20)

The discrepancy is only apparent, however, and should disappear when the issue of valence in its relation to affect is properly addressed. In emotion theory, the relations between valence and affect, and even the existence of valence,

are highly debated. With rare exceptions,¹ the question is ignored by theorists of somatic, agentive, or noetic feelings.² The proposal summarized in (2), however, posits that affect will result from valence (not the other way round). Section 7 aims to explain why affect depends on the dynamics of valence throughout the domain of feelings. These relations are modulated by the dynamic conditions that prevail in the contrast between expectancy and observation in a given domain. When observation and expectancy coincide with a predicted temporal pattern - with a small stake involved-, the corresponding feelings should not involve affect on top of valence. This is the case for the feelings of agentive success that are generated in routine actions. Hence (1) holds. When you predictably overcome a minor obstacle, you don't feel particularly thrilled. When special dynamic conditions obtain, however, (acceleration or deceleration in the rate of observed change, as compared with the expected rate of change), valence will be intensely felt, in terms of vividly positive or negative experiences. Scoring an ace in a tennis game, especially if it is a rare achievement for this player, elicits in him/her an intensive positive affect. Dynamic variations of this kind also apply to metacognition, where Archimedes' "Eureka" is affect-laden, while the felt ability to respond to a memory question in a laboratory is not.

Hence there may be affect-laden feelings beyond the domain of what is traditionally called "emotional" or "affective feelings". Reciprocally, one might suspect that in the latter domain, too, affect only appears beyond thresholds of positive or negative valence, with colder kinds of feelings occupying the lower end of the continuum.

2 Causal information: Explicit versus implicit

Iuliia Pliushch presents my view on the role of causal relations in feeling representations as fol-

lows: "Due to their non-conceptual monitoring nature, feelings do not convey, but merely approximate a causal relation between internal states and actions" ([this collection](#), p. 2). It may be useful to briefly comment on this summary, in order to clarify the aim of the passage where this question is discussed as follows:

Clearly, FS does not explicitly convey a causal relation between situation, somatic markers and subjective feeling. It carries this causal relation implicitly, however, as a consequence of the control architecture that produces feelings. In an emotional control loop, a perceived affordance causes (not: is represented as causing) its expressive evaluation through its specialized sensory feedback. Emotional awareness expresses this functional relation. ([Proust this collection](#), p. 11)

What is at stake is not the causal relation between internal states and actions, but rather the nature of the causal relation between, on the one hand, the agent's perceptual belief about an external situation ("there is a bear in front of me") and his/her own bodily changes (pounding heart, trembling legs, etc.). According to cognitivists, this causal relation is not only generating a specific emotion, or in my terms, a given feeling, as most theories would accept. It also constitutes in part the intentional content of the experience of fear, or more generally, of an emotional experience. What I object to here is that the *representational* structure of feelings is not *constituted* by a conceptual representation of the causal link between an external fact and observed bodily changes. The causal relations are, rather, implicitly represented in a felt affordance through the dynamic relations between the associated, embodied cues for location, valence and intensity and the type of affordance perceived. Perceiving a bear elicits a bear-affordance (i.e., a feeling of fear of this bear). Even though, from an external viewpoint, one might say that identifying an object as dangerous has caused a disposition to act in the agent, from the viewpoint of the engaged agent, no such judgment needs to be

¹ In particular Carver & Scheier (1990, 2001) for feelings of agentive success or failure, and Stepper & Strack (1993) for noetic feelings.

² For an interesting philosophical discussion of the nature of valence, see Prinz (2010), against Solomon's skeptical stance (2003).

formed because the representation of a given affordance includes the relevant “causal” information in its associative dynamic structure. As suggested by Pliushch, being evaluative, feelings predispose to act adaptively. A disposition to act, then, is associated with an affordance, and with the bodily markers for valence and intensity constituting this affordance.

3 Phenomenology of feelings: Background or foreground?

Should we construe the phenomenology of feelings – the presence of a bodily change – as being in the foreground or in the background of consciousness? The article under review briefly discusses this issue (Pliushch [this collection](#), pp. 2-3): A feeling tends to be more explicitly felt as bodily when making a bodily need salient (feeling tired, feeling a pain in the joints), plausibly because its function is to motivate bodily-directed action. Although in so-called “affective feelings” ¹ the bodily phenomenology tends to recede to the fringe of consciousness, there are cases, as Iuliia Pliushch notes correctly, where it occupies center stage – think of Proust’s report about his chest pain when learning that Madame de Guermantes just died.

It is debatable, however, that in such cases, the formal object of the feeling consists merely in the bodily changes, say, in heartbeat rate. For such states are part of an intensifying negative affordance: the loss of a friend. The notions of “meta-emotion” and “meta-feeling”, which are used by Pliushch to discuss the amplification of a feeling might be captured either in purely dynamic terms, or in a conceptual reconstruction of the situation at hand. This interesting issue, discussed in section 2.2 of Iuliia Pliushch’s comments, has connections with the notion of how concepts and feelings interact, and will be addressed in section 4.

4 Directedness

Iuliia Pliushch objects to my distinction between perceptions and feelings. The claim that “feelings do not have a direct sensory access to the world”, she says, relies on a meaning

of “direct” that is not compatible with the view defended by predictive coding theorists, where “directness is an absence of the evidentiary boundary” (Pliushch [this collection](#), p. 5). Being direct, then, if I understand this sentence correctly, means to lack independent evidence about the world of the kind that perception could bring. Although predictive coding offers a stimulating scheme for understanding mental function, it is open to interpretation and controversy. The functional hypothesis that perceiving and feeling are both indirect will appear highly counter-intuitive to many psychologists and philosophers.

As far as my article is concerned, I have defended the view that feelings are directly related to an opportunity, in the sense that they represent it in an immediate way, a view that has been defended by most affordance theorists. This is compatible with the claim that their informational pathway is derived from perception or memory. What may appear puzzling in my proposal is that an affordance is neither directly *perceived* nor *inferred*. It is directly *sensed*, which requires a different kind of ability. In section 5.1, I have proposed to distinguish associations from inferences, which is relevant to the present discussion. The kind of trigger for feelings are cues elicited in a currently active context, not inferences. These cues are delivered by sensory perception or by memory, but dealt with in a separate subsystem.

5 What are the relations between feelings and conceptual representations?

The comments in section 2.2 of my reviewer’s contribution are presented as an alternative approach to my own view, but I find myself in agreement with most of the claims, in particular with the remarks on p. 6 concerning the relations between feelings and conceptual representations. The main point concerns how one’s own goal, when acting, may influence the production of particular feelings. I discuss this issue at length in sections 5 and 6 of the article under review (Proust [this collection](#)), as well as in a recent publication devoted to action representa-

tions (Proust 2014). My position is captured by two claims. 1) Feelings – affordance sensings – can be, and indeed are usually triggered while performing a task that has been defined in conceptual terms. Cognitive affordances, in particular, are important relational properties that an agent needs to use when attempting to solve highly complex problems, for example when playing chess or looking for a mathematical proof. 2) The feeling episode, however, has an exclusively evaluative, non-conceptual content. I am aware that these two claims may easily be misunderstood. To disentangle the two, think of what agents mean to do: they mean to play chess according to the rules, or to prove a theorem. These goals, indeed, are conceptually represented, and depend on background beliefs and a sensitivity to epistemic norms such as truth and coherence, which presupposes in these particular cases an ability to represent beliefs as beliefs. Feelings of knowing, feelings of being right, and other affordance sensings are generated while the agents are conducting these higher-level forms of reasoning. They are dependent on the mental and neural activity which is thereby elicited. In other words, these feelings do not result from *a consideration* of the concepts involved, but from the dynamic features of the underlying processes. Hence, I would go farther than my reviewer, when she claims that noetic feelings are often elicited when concepts are automatically activated when forming a cognitive goal: they are also elicited when concepts are activated in a controlled way, e.g., in the process of planning what to do.

Should we conclude from this claim that heuristic processes are “contaminated by background knowledge” (Pliushch this collection, p. 6)? No. One should rather conclude that while the goal of a mental action is conceptually defined, the feelings entertained while acting are generated not by the concepts themselves, but by the dynamic characteristics of the processes underlying concept use. It is thus perfectly coherent to conclude that feelings have their own representational format that is not itself “infected” by concepts. A “theory of the task” is not a constituent of an affordance sensing, it is only a precondition for evaluating one’s ability in solving a task.

6 Serial versus dynamic properties of cognitive processes

My reviewer attributes to me a serial view of cognitive processes because I distinguish predictive from retrodictive evaluations of mental actions (Pliushch this collection, pp. 7-8). I do not think that this distinction commits me to serialism however. In my 2013 book, I propose that “a mind should primarily be seen as consisting of a hierarchy of control-and-monitoring loops, and their essentially dynamic interaction with the world, rather than as constituted by the successive states that emerge from this interaction”. Examples of how the dynamics at lower levels of representation can influence higher levels, and the converse, are discussed in chapters 11 and 12, where the case of schizophrenic delusions is analyzed. Hence, I have no problem with the view that low-level appraisal affects higher-level appraisals: these types of influences are part of what it is to have a hierarchy of control. This does not mean, however, that predictive appraisal and retrodictive appraisal should be conflated: they have a different evaluative function, and are based on different dynamic cues. This does not mean, either, that a concept-based judgment can easily influence an affordance-based appraisal. The difficulty of having a prolonged *strategic control* over one’s feelings (based on what one knows, as in the anagram experiment), originates in the different roles of associative cues and inferential relations between concepts in mental activity.³

Iuliia Pliushch is right, however, when observing that I stick to the distinction between feelings and their propositional re-description. From the viewpoint of action theory, this distinction corresponds to the contrast between reacting and acting strategically. I subscribe also to her remarks on p. 6, according to which goal representations might change affordance-sensings. The point is: how sustained is this change? A conceptual re-description tends to modify one’s representation of the context, and hence of one’s goals, which might either favor or re-

³ This point is developed in Proust (2014). A third form of action, habitual or routine action, is claimed to pertain to a second affordance-based system with its own agentic feelings of opportunity.

duce *further elicitation of feelings* (for example, by being ashamed of having felt anger), and even inhibit the influence of feelings on action. This is the case for the participants' epistemic decisions in phase 2 of the Anagram Experiment discussed in the section 5.2 of the article. Their ability to control their feelings, however, cannot resist time pressure and/or divided attention in phase 3.

On the view that I propose, feelings can only be sustainably modulated by having other feelings replace them. There are both automatic and strategic ways of enhancing one's feelings through other feelings (see [Proust 2014](#)). Feelings can easily be enhanced by enriching the associative representations constituting an affordance. Deliberately suppressing them, or reorienting them to new targets, however, is very difficult (as rejected lovers know all too well). The Confucian moral practices offer a very good example of a strategic attempt to train new moral feelings in followers (see [Reber 2013](#)). As Rolf Reber shows in his fascinating analysis of what he calls critical feelings, strategically redirecting one's feelings to new targets can only be performed by manipulating the fluency of one's own re-descriptions and conceptual rules for acting morally. In other terms, the agents need to be trained until they entertain feelings of ease of processing (i.e., feelings of fluency) when activating target concepts and inferences, rather than merely trying to immediately subsume their own initial feelings under critical concepts.

7 Self-deception and metacognition

Iuliia Pliushch finally makes an interesting suggestion: when self-deception occurs, the believer senses a metacognitive feeling of uneasiness, indicating that her underlying belief-forming process is faulty. This suggestion offers an account of the tension that arises while forming a belief on the basis of motivational, rather than evidential grounds. It would be wrong to interpret her proposal as the claim that finding faulty a belief, or a belief-forming process, involves an appraisal of the content of the belief, or of the kind of process that has been used to form it. As I understand her, Pliushch is rather claim-

ing, as psychologists and neuroscientists of metacognition do, that the mind is able to detect fault in the dynamical properties of the underlying processes. Pliushch argues further that, in contrast (she claims) with my own proposal, monitoring not only occurs "before or after a cognitive process, but also *during it*". There is no real conflict, however, about this claim. Presence of intermediate monitoring depends on the temporal extension of the mental action considered. When confronted with perceptual or memorial uncertainty, there is only control-based, mainly unconscious, intermediate monitoring; intermediate becomes prominent, however, in prolonged, effortful actions, such as problem solving ([Ackerman 2013](#)). I agree with Pliushch, however, that representing a mental action merely in terms of a starting and end points misrepresents the facts: it is based on a serial view that does not fit the dynamic character of metacognition (as already discussed in section 7 above). The evidence presented in [Proust \(2013\)](#) suggests that retrospective evaluation is based on the underlying dynamic of the *whole* action (the rate of accumulation in favor of a dominant response, as well as the dispersion of the neural responses), while predictive evaluation is based on the dynamics elicited by the command for this action, as compared with a stored standard (the complexity of the feedback used is addressed in [Koriat et al. 2006](#)). An epistemic evaluation, however, has two functions: stop the action, and encourage its continued performance, hence the role of polar valence in motivating action, which is reflected in the bi-partition of evaluations in two classes. This is in close agreement with how predictive coding, as any other theory of emotion and action, describes the facts.

Does predictive coding offer *new* insights on metacognition? The concept of "transition probabilities" mentioned by Pliushch, is shared by all theorists working on neural dynamics, as well by theorists of recurrent feedback; the concept of free-energy minimization, related to the minimization of surprise, seems *prima facie* to be consonant with [Rescorla & Wagner's \(1972\)](#) well established model of reinforcement learning. There is an internal connection

between free energy minimization and the evaluation of one's own uncertainty, because it is adaptive to predict one's chances of being incorrect, and hence avoid surprising failures. The concept of free energy, however, is no more equipped to provide any mechanistic account of brain function as any other evolutionary theory. "It is nothing more than the principle of least action applied to information theory", Friston recognizes (Friston et al. 2012). Indeed a prominent problem remains to be solved, concerning how priors vary as a function of task demands and of environmental statistics. Unpacking the principle across adaptive time-scales and survival contexts is indeed a complex future goal. Ways in which predictive coding might enrich the analysis of metacognition with new descriptive, operational tools or new functional explanations remain, then, to be specified.

Pliushch claims further that a first step in the proposed metacognitive theory of self-deception consists in recognizing that metacognitive feelings must be "extended to unconscious belief forming processes". If what is meant is that the dynamic properties that elicit feelings belong to such processes, there is universal agreement on this claim (see the so-called "cross-over principle" between unconscious heuristics and representations (including beliefs) and conscious feelings in Koriat 2000). What is meant, then, by the suggested "extension" is unclear. If what is meant, rather, is that the feelings themselves might be unconscious, this is a possibility that is taken seriously in studies of metaperception in blindsight patients (Reder & Schunn 1996). The very existence of such feelings complicates the phenomenologist's task. A second step is claimed to consist in "clarifying the representational content of tension". Although more detailed work needs to be done in order to better understand the contrast between perceptual and conceptual fluency, intuitivity is generally identified as a variety of what experimental psychologists call "feelings of fluency". One suggestion is that what creates feelings of tension or dysfluency in self-deception is not merely the representation that "the cognitive process violates some important goal representation", but rather, that it violates an implicit heuristic of

self-consistency, as discussed in Koriat (2012). Another suggestion is that tension has to do with the realization that the effort initially planned for a current task needs to be upgraded, which is a source of anxiety (Ackerman 2013). In summary: belief-forming processes are known to elicit metacognitive feelings. It remains to be shown how a metacognitive analysis of self-deception might enlighten philosophical and epistemological views about it. Self-deception is a good test case for making the point that conceptual-inferential processing also conveys non-conceptual information.

8 Serial versus dynamic properties of cognitive processes

As noted in the title of an article by Koriat et al. (2006), the relations between control and monitoring in the production of metacognitive feelings are very "intricate". Iuliia Pliushch's insightful comments have initiated what I hope to be a useful clarification of another aspect of feelings (whether metacognitive or not): their relations with propositional thoughts. Feelings elicited by tasks that are conceptually characterized do not become *ipso facto* conceptually penetrable: this difficult, unintuitive claim is often misunderstood and resisted for wrong reasons, which does not mean that it would resist any reason! The objection related to serialism was odd, given my own interest in the dynamic properties of the mental processes as offering a source of information that stable propositional properties of mental contents cannot provide. Once prediction and post-evaluation are identified as two major functions in metacognition, it is indeed important to emphasize that metacognitive processes of each kind are dynamic, and rely on various types of re-afferent feedback. Epistemic decisions, however, once made, are discontinuous by design, which turns the pre-decisional confidence level into a final evaluation that triggers or inhibits the corresponding action. Hence, a contrast must be maintained between how to select a goal and determine the level of effort needed to achieve it (i.e., a control command), on the one hand, and monitoring progress toward the goal, on the other hand.

Each form of metacognition elicits feelings. This does not mean that the two functions need to be serially executed: for long, effortful tasks, agents need to frequently revise their level of effort and of success expectancy, by monitoring over time their progress through associated heuristics and feelings.

References

- Ackerman, R. (2013). The diminishing criterion model for meta-cognitive regulation of time investment. *Journal of Experimental Psychology: General*, 143 (3), 1349-1368. [10.1037/a0035098](https://doi.org/10.1037/a0035098)
- Carver, C. S. & Scheier, M. F. (1990). Origins and functions of positive and negative affect: A control-process view. *Psychological review*, 97 (1), 19-35. [10.1037/0033-295X.97.1.19](https://doi.org/10.1037/0033-295X.97.1.19)
- (1990). Origins and functions of positive and negative affect: A control-process view. *Psychological review*, 97 (1), 19-35.
- (2001). *On the self-regulation of behavior*. Cambridge, UK: Cambridge University Press.
- Friston, K. J., Thornton, C. & Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in Psychology*, 3 (130). [10.3389/fpsyg.2012.00130](https://doi.org/10.3389/fpsyg.2012.00130)
- Koriat, A. (2000). The feeling of knowing: Some metatheoretical Implications for Consciousness and Control. *Consciousness and Cognition*, 9 (2), 149-171. [10.1006/ccog.2000.0433](https://doi.org/10.1006/ccog.2000.0433)
- (2012). The self-consistency model of subjective confidence. *Psychological Review*, 119, 80-113.
- Koriat, A., Ma'ayan, H. & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General*, 135 (1), 36-69. [10.1037/0096-3445.135.1.36](https://doi.org/10.1037/0096-3445.135.1.36)
- Pliushch, I. (2015). The extension of the indicator-function of feelings. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Prinz, J. (2010). For valence. *Emotion Review*, 2 (1), 5-13. [10.1177/1754073909345546](https://doi.org/10.1177/1754073909345546)
- Proust, J. (2013). *The philosophy of metacognition. Mental agency and self-awareness*. Oxford, UK: Oxford University Press.
- (2014). Time and action: Impulsivity, habit, strategy. *Review of Philosophy and Psychology*. [10.1007/s13164-014-0224-1](https://doi.org/10.1007/s13164-014-0224-1)
- (2015). The representational structure of feelings. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Reber, R. (2013). Critical feeling. In C. Unkelbach & R. Greifeneder (Eds.) *The experience of thinking* (pp. 173-189). Hove, UK: Psychology Press.
- Reder, L. M. & Schunn, C. D. (1996). Metacognition does not imply awareness: Strategy choice is governed by implicit learning and memory. In L. M. Reder (Ed.) *Implicit memory and metacognition* (pp. 45-78). Mahwah, NJ: Lawrence Erlbaum Ass.
- Rescorla, R. A. & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.) *Classical conditioning II: Current research and theory* (pp. 64-99). New York, NY: Appleton Century Crofts.
- Solomon, R. C. (2003). Against valence ('positive' and 'negative' emotions). In R. C. Solomon (Ed.) *Not passion's slave* (pp. 162-177). Oxford, UK: Oxford University Press.
- Stepper, S. & Strack, F. (1993). Proprioceptive determinants of emotional and nonemotional feelings. *Journal of Personality and Social Psychology*, 64 (2), 211-220. [10.1037/0022-3514.64.2.211](https://doi.org/10.1037/0022-3514.64.2.211)

The Avatars in the Machine

Dreaming as a Simulation of Social Reality

Antti Revonsuo, Jarno Tuominen & Katja Valli

The idea that dreaming is a simulation of the waking world is currently becoming a far more widely shared and accepted view among dream researchers. Several philosophers, psychologists, and neuroscientists have recently characterized dreaming in terms of virtual reality, immersive spatiotemporal simulation, or realistic and useful world simulation. Thus, the conception of dreaming as a simulated world now unifies definitions of the basic nature of dreaming within dream and consciousness research. This novel concept of dreaming has consequently led to the idea that social interactions in dreams, known to be a universal and abundant feature of human dream content, can best be characterized as a simulation of human social reality, simulating the social skills, bonds, interactions, and networks that we engage in during our waking lives. Yet this tempting idea has never before been formulated into a clear and empirically testable theory of dreaming. Here we show that a testable Social Simulation Theory (SST) of dreaming can be formulated, from which empirical predictions can be derived. Some of the predictions can gain initial support by relying on already existing data in the literature, but many more remain to be tested by further research. We argue that the SST should be tested by directly contrasting its predictions with the major competing theories on the nature and function of dreaming, such as the Continuity Hypothesis (CH) and the Threat Simulation Theory (TST). These three major theories of dreaming make differing predictions as to the quality and the quantity of social simulations in dreams. We will outline the first steps towards a theory-and-hypothesis-driven research program in dream research that treats dreaming as a simulated world in general and as a social simulation in particular. By following this research program it will be possible to find out whether dreaming is a relatively unselective and thus probably non-functional simulation of the waking world (CH), a simulation primarily specialized in the simulation of dangerous and threatening events that present important challenges for our survival and prosperity (TST), or whether it is a simulation primarily specialized in training the social skills and bonds most important for us humans as a social species (SST). Whatever the evidence for or against the specific theories turn out to be, in any case the conception of dreaming as a simulated world has already proved to be a fruitful theoretical approach to understanding the nature of dreaming and consciousness.

Keywords

Altered state of consciousness | Avatar | Consciousness | Continuity hypothesis | Dreaming | Evolutionary psychology | Inclusive fitness | Kin selection theory | Need to belong | Practise and preparation hypothesis | Reciprocal altruism theory | Simulation | Social brain hypothesis | Social mapping hypothesis | Social simulation theory | Sociometer theory | Strengthening hypothesis | The dream self | The inclusive fitness theory | Threat simulation theory | Virtual reality | Virtual reality metaphor

Authors

[Antti Revonsuo](#)
antti.revonsuo@utu.fi
Högskolan i Skövde
Skövde, Sweden
Turun yliopisto
Turku, Finland

[Jarno Tuominen](#)
jarno.tuominen@utu.fi
Turun yliopisto
Turku, Finland

[Katja Valli](#)
katval@utu.fi
Turun yliopisto
Turku, Finland
Högskolan i Skövde
Skövde, Sweden

Commentator

[Martin Dresler](#)
martin.dresler@donders.ru.nl
Radboud Universiteit Medical Center
Nijmegen, Netherlands

Editors

[Thomas Metzinger](#)
metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Introduction

There may be no Cartesian ghosts residing within the machinery of the brain, but still, something rather peculiar is going on in there, especially during the darkest hours of the night. As we sleep and our bodies cease to interact behaviourally with the surrounding physical world, our conscious experiences do not entirely disappear. On the contrary, during sleep we often find ourselves embodied and immersed in an experiential reality, an altered state of consciousness called dreaming. The Dream Self—the character with which we identify ourselves in the dream world, and from whose embodied perspective the dream world is experienced—is who *I* am in the dream world (Revonsuo 2005).

But we are not alone in this alternative reality—there are other apparently living, intelligent beings present, who seem to share this reality with us. We see and interact with realistic human characters in our dreams. Their behaviour and their very existence in the dream world seem to be autonomous. The dream people who I encounter within the dream seem to go about their own business: I cannot predict or control what they will say or do. Yet, they, too, are somehow produced by my own dreaming brain.

On the one hand, dreaming is a solipsistic experience: when we dream, we dream alone, and outsiders have no way of participating in our dream. Yet on the other hand, dreaming is an intensely *social* experience, even if the social contacts and interactions in the dream world are merely virtual. In this paper, we will explore the idea that dreaming is a *simulated* world, but not only a simulation of the *physical* world. It is equally or perhaps even more importantly a simulation of the *social* world. We will proceed in the following way:

First, we will argue that a remarkable convergence has gradually emerged in theories about the nature of dreaming. The field used to be a disunified battleground of directly opposing views on what dreams are, how exactly the concept of “dreaming” should be defined, and on the proper level of description and explanation for dreaming. Recently, the field has con-

verged towards a more unified understanding of the basic nature of dreams. A widely shared conceptualization of dreaming now depicts it as the *simulation* of waking reality. We will briefly describe how this theoretical shift has taken place and where we currently are in the theoretical definition of dreaming. This theoretical development has paved the way for understanding the *social* nature of dreams in terms of social simulation.

Second, we will explore the nature of social dream simulation in more detail. In what sense can dreaming be taken as a simulation of our human *social* reality? How much and what types of social perception and interaction occur in dreams? This question can be broken down into a number of more detailed questions. We will try to answer some of these questions based on the already existing knowledge and empirical evidence about the social nature of dreams. Furthermore, we will try to formulate more clearly the questions that cannot yet be answered empirically due to the lack of appropriate data.

Third, we will review hypotheses that already address the question of the social nature of dreams or assign a social simulation function for dreams. Finally, we will outline some basic ideas of a Social Simulation Theory (SST) of dreaming that might offer some explanations for the social nature of dreams, or at least might produce well-defined, testable research questions concerning the possible *functions* of social dream simulations.

To describe and explain the social nature of dreams as social simulation, concepts borrowed from virtual reality technology may be applied, in this case to the social aspects of dreaming. One of these concepts is the notion of “avatar”: *A simulated virtual human character* who plays the role of a corresponding real human within a virtual reality. If dreams are virtual realities in the brain (Revonsuo 1995), then we ourselves within the dream world are avatars, and we interact with other avatars inside the simulated reality. Somehow, the dreaming brain is capable of creating credible, autonomous human simulations out of neural

activities in the sleeping brain. A theory of dreaming as a social simulation should predict what kind of avatars are represented in our dreams, what types of interactions we engage in with them, and in particular, *why* it would be useful to simulate such avatars and interactions in our dreams—what functions, if any, do they serve for us.

2 Consciousness as reality-modeling and world-simulation

Dreaming is the most universal and most regularly occurring, as well as a perfectly natural and physiological (as opposed to pathological), altered state of consciousness. Thus, any plausible (empirical or philosophical) theory of consciousness should also describe and explain dreaming as a major state of consciousness. Most theories of consciousness, however, do not consider dreaming at all or at least do not discuss the results of dream research in any detail (Revonsuo 2006).

Dreaming presents a particularly difficult challenge for externalist, embodied, and enactive types of theories of consciousness.¹ They all anchor the existence and nature of consciousness to something in the world external to the brain, or to some kind of brain-world relations that, at least partly, reside outside the brain. By contrast, the empirical evidence from dream research shows that full-blown, complex subjective experiences similar with or identical to experiences during wakefulness (e.g., Rechtschaffen & Buchignani 1992), regularly and universally happen during rapid eye movement (REM) sleep. The conscious experiences we have during dreaming are isolated from behavioural and perceptual interactions with the environment, which refutes any theory that states that organism-environment interaction or other external relationships are constitutive of the existence of consciousness (Revonsuo 2006).

A few theories of consciousness have, however, taken dreaming as a central starting point in their conceptualization and explanation of

consciousness. When dreaming is taken seriously, ideas about the nature of consciousness tend to converge on internalist theories of consciousness that take consciousness and dreaming to be varieties of the same internal phenomenon, whose main function is to simulate reality.

One of the earliest attempts to conceptualize both waking consciousness and dreaming as the expressions of the same internally-activated neural mechanism, only differently stimulated, was put forward by Llinás & Paré in 1991:

[C]onsciousness is an intrinsic property arising from the expression of existing dispositions of the brain to be active in certain ways. It is a close kin to dreaming, where sensory input by constraining the intrinsic functional states specifies, rather than informs, the brain of those properties of external reality that are important for survival. [...] That consciousness is generated intrinsically is not difficult to understand when one considers the completeness of the sensory representations in our dreams. (1991, p. 531)

The argument by Llinás & Paré (1991) was mostly based on considerations of the shared neurophysiological mechanisms (in the thalamo-cortical system) that could act as the final common path for both dreaming and waking consciousness. Binding information together within this system intrinsically generates consciousness (“It binds, therefore I am”, Llinás 2001, p. 261); but only during wakefulness is consciousness modulated by sensory-perceptual information—in this model, wakefulness can be seen as a dream-like state (Llinás & Ribary 1994).

Although the idea that dreaming *simulates* waking consciousness was implicit in this neuroscientific theory, Llinás & Paré (1991) did not consider the phenomenology of dreaming and consciousness in any detail. Theoretical approaches characterizing the nature of dreaming as simulation, based on a combination of philosophical arguments and empirical facts about dreaming, started to emerge during the 1990s. In Revonsuo (1995) the idea was put forward

¹ The same criticism may to some extent also apply to representationalist theories of consciousness and dreaming, depending on which externalist or internalist version of representationalism the theory is committed to.

that consciousness in general and dreaming in particular may best be characterized as a virtual reality in the brain, or a model of the world that places a (virtual) self in the centre of a (virtual) world. All experiences are virtual in the sense that they are world-models rather than the external physical world somehow directly apprehended. While the causal chains that modulate the virtual reality are different during wakefulness and dreaming, the virtual world is ontologically the same biological phenomenon: the *phenomenal level of organization* in the brain (Revonsuo 1995). All experiences are, according to this view, in their intrinsic phenomenal character, no different from dreams.

Metzinger (2003) took this line of thought further and analysed dreams as complex, multimodal, sequentially organized models of the world that satisfy several important constraints of consciousness. Dreams activate a *global* model of the world (globality), they integrate this model into a *window of presence* (presentationality), and this model is *transparent* to the experiencing subject, who takes it to be a real world and not a mere model of the world (transparency) (see also Windt & Metzinger 2007).

In *Inner Presence* Revonsuo (2006) presented a lengthy analysis and defence of the idea that dreams are internal virtual realities, or *world-simulations*, and argued that consciousness in general would be best described and explained by treating dreaming as a paradigmatic model system for consciousness. The world-simulation contains the *virtual self* and its *sense of presence* in the centre of the simulation. The virtual self is perceptually surrounded by the *virtual place*; the virtual place in turn contains multiple perceptual contents in the form of animate and inanimate *virtual objects*, including human characters. The virtual objects are bound together from phenomenal features like color, shape, and motion, but this binding in dreams does not always work coherently, thereby resulting in bizarre feature combinations and incongruous or discontinuous objects and persons in dreams (Revonsuo 2006).

Recently, Windt (2010) has formulated a definition of dreams that stems from similar ba-

sic ideas. Windt's definition aims to capture the minimal set of phenomenological features that an experience during sleep should have in order to count as a "dream" (as opposed to other types of sleep mentation). This definition, although not explicitly applying the concept of "simulation", is consistent with the world-simulation model of dreaming. According to Windt, dreams are Immersive Spatiotemporal Hallucinations (ISTH): there is a sense of spatial and temporal presence in dreams; there is a hallucinatory scene organized around a first-person perspective, and there is a sense of "now", along with temporal duration. The core feature of a dream experience is, in Windt's ISTH, *the sense of immersion or presence in a spatiotemporal frame of reference*. Thus, Windt's ISTH, as well as Metzinger and Revonsuo's earlier definitions, all involve similar ideas of dreams as involving an immersive presence of a virtual self in a virtual, spatiotemporally organized world-model or simulation.

3 Dreaming as simulation: Converging definitions from dream research

Within empirical dream research, definitions of dreaming have been highly variable and often motivated by underlying theoretical background assumptions held by the theorist. Thus, the pure description of the *explanandum*, which should come first in any scientific inquiry, has perhaps been biased by a pre-existing theory as to what might count as the *explanans*—the entities, processes, and concepts that are supposed to explain the phenomenon. We will only briefly mention three approaches to defining (and explaining) dreams in the recent history of dream research, where the definition and description of the data seem to have been theoretically motivated.

The field of dream research was, in the 1970–1990s, a theoretically disunified field. The deep disagreements over finding a definition of "dreaming" that would be acceptable across the field were noted by Nielsen (2000, p. 853)

[T]here is currently no widely accepted or standardized definition of dreaming.

as well as by Hobson et al. (2000, p. 1019):

[...T]here is no clearly agreed upon definition of what a dream is [...] and we are not even close to agreement.

Hobson's (1988, 1997, 2001) own definition of dreaming is (or at least was in his earlier writings) a list of some features of dream experience. According to him, a dream is mentation during sleep that has most of the following features: *hallucination, delusion, narrative structure, hyperemotionality, and bizarreness*. This definition may be (and was) criticized as including only paradigmatic late-night REM dreams that are spontaneously remembered and on which our everyday stereotype of what dreams are like is based. This bias in the definition towards REM dreams might be seen to reflect the underlying theoretical idea or commitment, obvious in Hobson's earlier theories, that dream phenomenology should be (reductively) explained by referring to the features of REM neurophysiology.

The opposing, cognitive-psychological view of the 1980s and 1990s conceptualized dreaming as a cognitive process that should be explained at the cognitive-psychological level (Foulkes 1985). References to the neurophysiological level were unnecessary. In that time and in the spirit of functionalism and classical cognitive science, the cognitive levels of description and explanation were in general seen to be completely independent of implementation levels, such as neurophysiology. Furthermore, dreaming was thought to occur in every stage of sleep, not only REM sleep, and rather than being full of bizarreness was mostly a credible replica of the waking world. Thus, according to the cognitive approach, an explanation of dreaming cannot be based on neurophysiological mechanisms in general, or for REM sleep on neurophysiology in particular. The explanation should be given at cognitive levels rather than neurobiological ones. Interestingly, it was probably Foulkes (1985) who first characterized dreams in terms of the idea and the concept of simulation. In 1985 he described dreams as *credible world analogs*, an organized form of

consciousness that *simulates* what life is like in a nearly perfect manner.

A third theoretical definition of dreaming came from clinical dream research, and reflected the long and widespread idea in clinical psychology that dreams restore our emotional balance and have a psychotherapeutic function. Hartmann formulated this definition of dreaming most clearly, when he said that "Dreaming, like therapy, is the making of connections in a safe place" (1996, p. 13).

During recent years in dream research, the concept of simulation has become a widely accepted way of characterizing and defining dreaming, as well as a way of formulating theoretical ideas about the potential functions of dreaming. Thus, the idea that *dreaming is a multimodal, complex, dynamic world-simulation in consciousness during sleep*, may be a type of conception and definition of dreaming that many if not most dream researchers are ready to accept (Nielsen 2010). The various contents of dreams—their events and objects and characters—can be taken to be simulations of their real-world counterparts.

Taking Foulkes's idea of dreams as credible world analogs and as the simulation of what life is like as a starting point for defining dreaming, Revonsuo (1995) formulated the Virtual Reality metaphor and later the TST (Threat Simulation Theory) of the evolutionary function of dreaming. This theory is built on two background assumptions, the first of which is precisely the definition of dreaming as "an organized simulation of the perceptual world" (Revonsuo 2000, p. 883). An additional, more specific assumption of this theory is that dream experience is *specialized* in particular in the simulation of *threatening* events: it tends to select and include various types of dangerous enemies and events and then simulates what it is like to perceive and recognize them (simulation of threat perception) as well as how to react and behaviourally respond to them (simulation of threat avoidance behaviours and strategies). Threat simulations appear in a paradigmatic and powerful form especially in nightmares, bad dreams, and post-traumatic dreams, but are also abundant in many other types of dreams

such as everyday dreams, recurrent dreams, and in various parasomnias such as RBD (REM-Sleep Behaviour Disorder).

Domhoff (2007), who represents a similar psychological and content-analysis approach to dream research as Foulkes (1985), also characterizes dreams as mostly realistic and reasonable *simulations* of waking life. By emphasizing that, according to convincing empirical data from content-analysis studies of dreams, dream simulations are mostly *realistic* rather than overly bizarre and hyperemotional, Domhoff argues against the Hobsonian definition of dreaming as being full of bizarre contents.

Still, despite their disagreements, both camps now seem to accept the notion of *simulation* as a valid description of the core nature of dreaming. Hobson, in his new *protoconsciousness* theory of dreaming and REM sleep (2009), uses the concept of simulation to characterize the root phenomenon, protoconsciousness, from which both our waking and dreaming consciousness arise. According to Hobson, protoconsciousness is the simulated experiential reality or a virtual reality model of the world that the developing brain turns on during REM sleep even before birth, to prepare the conscious brain to simulate the external reality that it will encounter through the senses after birth. This model of the world is genetic, innate, and a human universal. Protoconsciousness acts as the template on which both waking and dreaming consciousness are built after birth. Thus, according to this theory, protoconscious dream consciousness—a very basic form of an internally simulated world—comes into being prior to waking consciousness, and is causally necessary for waking consciousness. As Hobson (2011, p. 30) puts it: “I REM, therefore I will be”. According to Hobson & Friston (2012), *predictive coding* is an underlying mechanism in the brain that produces predictive simulations of the world. Therefore, dreaming may also function as a preparatory simulation of the waking world; thus their idea is closely related to the other simulation-theories of dreaming (Hobson & Friston 2012).

In conclusion, while there still are disagreements about many details of dream con-

tent and function, there seems to be relatively widespread agreement that the definition of dreaming includes the idea of “simulation” of the waking world. The use of the concept of “simulation” to characterize dreaming has recently gained wide acceptance in the field. The simulation is variously characterized as the simulation of waking life, of waking reality, or of waking consciousness, and variously called by different authors a realistic world-simulation, a virtual reality, an immersive spatiotemporal model of the world, and so on—but despite the somewhat varying terminology, the different terms seem to describe the same basic idea. This conceptual unification is a significant step forward in the theoretical description and explanation of dreams. It paves the way for a more unified theory of dreaming.

4 The simulation of social reality in dreams

Dreaming not only places us into an immersive (virtual) physical reality, but also immerses us into a (virtual) *social* reality: in dreams we are surrounded by close friends and family members, schoolmates, teachers and students, spouses, romantic partners, old crushes, colleagues and bosses, celebrities, politicians, acquaintances, strangers, and mobs as well as monsters and other fictitious characters from movies and video games. All are there in dream simulation with us as simulated characters—avatars—and we interact with these avatars in multiple ways: we perceive, recognize, and semantically classify them, we communicate and talk with them, we collaborate with them, help them, criticize them, fight them, escape them, fear them, and love them. At least intuitively, there is no doubt that in our dreams, we live rich and colourful social lives, even if only simulated ones.

If dreaming in general can be defined as a simulated world, the question arises whether the concept of “simulation” can also be usefully applied to describe the social reality of dreams. The first task for a theory that takes the concept of simulation seriously is to simply *describe* the social contents of dreams as simula-

tions of human social reality. The descriptive questions can be formulated in more detail along the following lines:

1. What kind of social perception, social interaction, and social behaviours are simulated in dreams?
2. How frequently are different kinds of social perception, interaction, and behaviour simulated in dreams? How much variation is there in the frequency of different social simulations as a function of gender, age, culture, and as a function of the quality and quantity of social interactions during waking life?

It is possible to find answers to many of the above descriptive questions from the already-existing dream research literature where various aspects of the social contents of dreams have been reported, even if they have not been conceptualized as social simulations. In what follows, we will first briefly review some of the major findings in the literature that describe the quality and the quantity of social simulation in dreams. Once we have detailed empirical descriptions of the quality and quantity of social simulations in dreams, we may seek explanatory theories and testable hypotheses that could account for why we have social simulation in dreams.

4.1 Evidence for simulation of social perception in dreams

From the already existing literature, it is possible to find statistics that describe the quality and quantity of social simulations in dreams. However, the theoretical concept of “social simulation” is rarely used in dream research literature for interpreting the descriptive results. Here, we will briefly summarize only some of the major findings.

The minimal criterion for a dream to count as a social simulation is that the Dream Self is not alone in the dream but in the presence of at least some other animate character or characters. In less than 5% of dreams is the dreamer alone (Domhoff 1996); thus, on this minimal criterion, dreaming seems to consist-

ently simulate social reality. The other animate characters simulated in dreams are predominantly human (normative finding in adults is about 95% human, 5% animal), but the proportion of animal characters varies in different cultures and age groups, being highest (up to 30–40%) in young children and in adults in hunter-gatherer societies (Domhoff 1996; Revonsuo 2000). As human characters are reported in almost all dreams, and typically there are two to four non-self characters in a dream (Nielsen & Lara-Carrasco 2007), the presence of simulated human characters must be perceptually detected and registered in the dream by the dreamer. Thus, during dreaming, *our neurocognitive mechanisms constantly simulate social perception*.

The minimal form of social perception is to *detect or register the presence* of some human character. A more sophisticated form is the perceptual *recognition* and *identification* of the human characters who are present, first in terms of some basic perceptual and semantic categories (male/female; familiar/stranger), and then in terms of more detailed semantic and autobiographical information about the precise identity and name of the person. According to the Hall and Van de Castle norms, about 90% of simulated human characters have sufficiently definite characteristics to be semantically categorized, for example as male or female, or as familiar or unfamiliar (Domhoff 1996). Thus, social recognition and identification mechanisms are highly engaged in almost all cases of social perception in dreams. The dreamer knows, both during the dream and afterwards when reporting it, whether the simulated characters present in the dream are (or were) male or female, familiar or strange, friend or family; and in most cases, the familiar characters are identified as particular persons from real life.

Typically, a slight majority of dream characters are avatars for familiar persons, although there are well-established gender differences (Domhoff 1996) that might, however, partly depend on the gender distribution encountered in the real-world social environment (Paul & Schredl 2012). In a sample of five hundred REM dreams (Strauch & Meier 1996) familiar people

(friends, acquaintances, and relatives) were simulated most frequently (44% of all characters), strangers represented about 25% of dream characters, and undefined people about 19%. In most dreams, both familiar and unfamiliar people were simulated, but in 30% only strangers and in 20% only familiar people appeared. The mixture of familiar and unfamiliar people was true also at the individual level—there were no participants who would have simulated only strangers or only familiar people in their dreams.

For the most part, the human avatars in the dream world are quite *realistic* simulations of their waking counterparts. The degree of realism, however, is difficult to express with accuracy by any single measure or quantity, as there are several features of human characters that may independently vary along the dimension of realism (Revonsuo & Tarkko 2002). The opposite pole for realism is called *bizarreness*, which in dream research refers to deviation from the corresponding entity in waking life.

If any kind and degree of deviation from a waking counterpart is counted as a bizarre feature of a simulated person, then over half of the simulated humans in dreams (over 60% according to Kahn et al. 2002; 53% according to Revonsuo & Tarkko 2002) are not perfectly realistic simulations. In contrast to other dream characters the Dream Self is rarely distorted in any way (Revonsuo & Salmivalli 1995). Revonsuo & Tarkko (2002) also found that in the vast majority of cases (around 90% of dream characters), non-self dream characters are *perceptually* entirely realistic—they *look* the same as their counterparts look in real life. Where they deviate from their counterparts is most often their verbal and nonverbal behaviour. Thus, although the perceptual simulation of human characters is nearly flawless in dreams, the simulation of expected or predicted *behaviours* deviate from waking norms relatively often, though still at least a slight majority of behaviours by dream characters are no different from waking life.

Dream characters are also spatially and temporally quite stable and continuous within the dream, although transformations and discontinuities sometimes do happen (Nielsen &

Lara-Carrasco 2007). A simulated person sometimes appears from nowhere, is magically transformed into someone else, or suddenly disappears without a trace. But these kind of discontinuous features account for less than 5% of dream character features (Revonsuo & Salmivalli 1995; see also Revonsuo & Tarkko 2002).

By contrast, the behaviours expressed by dream characters are relatively often to some extent odd or unpredictable. Thus, the simulated social reality in dreams is *less predictable* than the corresponding social reality during wakefulness. However, it is unclear how this unpredictability should be interpreted: does it simply reflect the difficulty (and consequently failure) of simulating complex human behaviours and interactions realistically by the dreaming brain, or is there some other more functional explanation as to why the avatars in our dreams tend to behave in more erratic ways compared to their waking-life counterparts? We will come back to this question when we consider the possible functions of social simulation in dreams.

4.2 Evidence for simulation of social interactions in dreams

The Dream Self and other dream characters are simulated in almost all dreams, but how often are they engaged in mutual social interactions? According to Strauch & Meier's (1996) data (140 REM dreams in which a Dream Self was present and had an active role), in nearly 50% of these dreams the Dream Self and characters interacted, in an additional 20% they acted together, and in 20% they acted independently of each other. In the rest, the Dream Self acted alone. Thus, social interaction or acting together is typically simulated in dreams where the Dream Self is present together with some other dream characters. When social interaction takes place, there is almost always verbal communication or conversation between the Dream Self and the other characters, which tends to be focused on concrete topics (Strauch & Meier 1996), and it is understandable and something that would be sayable in waking life (Heynick 1993).

The more detailed nature of social interactions has typically been categorized in terms of “friendly” and “aggressive” interactions. Friendly interactions are on average found in about 40% of dreams, whereas aggressive interactions are somewhat more common, and occur in about 45% of dreams in a normative sample (Domhoff 1996). Strauch and Meier, however, point out that in their sample, neutral interactions were also common, and only about half of the social interactions in their sample could be classified as particularly friendly or aggressive. The third category of social interactions that has typically been quantified in dream reports is sexual interactions, but they occur at a very low frequency—in Strauch & Meier’s (1996) laboratory data, in less than 1% of REM dreams, and in the normative Hall and Van de Castle (Domhoff 1996) data, in 4% of women’s and in 12% of men’s dreams collected in a home setting.

In sum, the simulation of dream characters occurs very frequently, the characters are perceived and recognized by the Dream Self, and the Dream Self actively participates in communication, social interaction, and joint actions with the characters. The simulated characters are also for the most part realistic, stable, and represent a variety of different kinds of people. Their behaviours, however, may sometimes be unusual or inappropriate, and not exactly what we would have expected from their counterparts in real life. The tone of the interactions may be neutral, friendly, or aggressive.

When this evidence is taken together, we may conclude that dreaming simulates a rich, variable, realistic, and concrete, but somewhat unpredictable social reality, inhabited by a mixture of familiar, unfamiliar, and undefined people. Therefore, we have solid grounds to state that dreaming *is*, among other things, definitely a social simulation. If this is a universal and ubiquitous feature of dreaming, what kind of theory could explain it? *Why* does dreaming simulate social reality at all? It is by no means self-evident that this should be the case. Dreaming could as well be only a simulation of some basic features of the physical world: space, time, objects, events, and the perception of and bodily interaction with the physical world. Or it

could be a simulation of thought processes, a thinking-through of our problems, or of our emotional states and concerns. Moreover, simulation of physical objects and their behaviour, or a replay of thinking and emotions, would probably be a simpler task for the brain than the simulation of a complex social world. Simulation of human bodies and faces and interactive behaviours such as conversations seems to require a lot of energy and computing power—these are very complex phenomena to simulate realistically. Thus, *why* does the sleeping brain simulate social situations in such an intense and invariant manner? Is there any convincing theoretical answer to be found to this question?

5 The continuity hypothesis and social simulation theories of dreaming

There are, of course, countless theories of dreaming. Some have explicitly considered the role of social interactions in dreams, while others make more general statements about dream content. One of the latter is the Continuity Hypothesis (CH), which states that dreams *reflect* waking life experiences (Schredl & Hofmann 2003) or, more specifically, that our waking concerns, thoughts, and experiences have a *causal influence* on subsequent dream content. Thus, if certain types of social contacts or interactions become more frequent (or less frequent) in waking life, their simulation in dreams becomes correspondingly more (or less) frequent.

This general principle seems to hold in many cases. For example, in hunter-gatherer societies, where people perceive and interact with wild animals on a daily basis, the proportion of animal characters remains high (as it is in children’s dreams across cultures), whereas in highly industrialized societies, the animal percentage decreases dramatically from childhood to adulthood. But the CH merely restates this empirical relationship; it cannot answer the theoretical question of *why* in young children’s dreams the proportion of animal characters *is high to begin with*. TST (Revonsuo 2000) has attempted to answer this question by referring not to personal experiences in waking life, but to a universal bias that is built into the default

values of dream content during human evolutionary history.

The CH, even if on the right track in many cases, is too vague and general as a theoretical explanation of the details of dream content. It does not predict in any detail how and why the causal relationship between waking and dreaming works. It also does not specify in any detail what counts as a “continuity” and what would count as a “discontinuity” between waking life experiences and dream simulations of the same. If something happens in waking life *how closely similar* will the dream simulation be to its waking origin, *when* will the same (or a similar) content appear in dreams, *how frequently and for how long* will it be incorporated into dreams, and so on? These questions have been studied under the concepts of day residue (Freud 1950) and the dream lag effect (Nielsen & Powell 1989). The CH takes almost any similarity between waking life and dream life as a confirmation of the continuity hypothesis. But “similarity” as a relationship between two phenomena is undefined, ambiguous, and vague. Something that in one respect is similar to its waking counterpart is in another respect dissimilar from it; thus it can be interpreted as either continuous or as discontinuous with waking life. Obviously, if the very same evidence could be counted as either supporting or disconfirming a theory, there is something wrong with how the theory is formulated.²

As long as the CH remains vaguely formulated, almost anything can be counted as its support. If the hypothesis does not specify in any detail the potential empirical observations after which its predictions would be falsified, it is not an empirically testable theory. Unless it is formulated in a much more specific manner, so that risky, exact predictions can be derived from it, its explanatory power remains correspondingly weak. In one study where more precise predictions from CH were derived, the CH was found not to be valid as a general rule concerning how often different everyday activities are reflected in dreams (Schredl & Hofmann 2003).

Perhaps a more precise prediction that could be derived from CH can be formulated in the following way: according to CH, dreams represent a random sample of recent waking experiences (or a random sample of their memory representations). The quantities of different types of contents in dreams will therefore passively reflect the proportion of their occurrence in waking life in the recent past (or the memory representations of waking life). If CH is formulated in this manner, as a prediction of random sampling and passive mirroring of recent waking life, then any systematic deviation from a random sample of waking contents (or memories thereof) would count as evidence against the CH. A deviation from passive mirroring of waking life would suggest that some kind of *selective* mechanism is at work. An *active selection* bias of particular contents to be either included in dreams or to be left out would be expected to result in a disproportionately exaggerated or diminished frequency of that content in dreams as compared with waking life. This kind of formulation of the predictions of CH makes it a testable theory.

Some more specific suggestions about dreaming as social simulation have been put forward in the literature. Brereton’s (2000) Social Mapping Hypothesis suggests that dreaming simulates, among other things, the awareness of other persons (social perception) and their internal mental states (mentalizing or theory of mind-abilities). This theory proceeds from an evolutionary standpoint, and considers dreaming as a rehearsal ground for emotional and perceptual abilities related to the mapping of the body image of the self into an emotionally-salient social space. Others have also hypothesized that our mindreading abilities could potentially be a target of simulated social perception in dreams (Kahn & Hobson 2005; McNamara et al. 2007). Moreover, Nielsen & Germain (2000) have suggested that dreaming might simulate attachment relationships and interpersonal bonds in ways that would maintain their adaptive significance even today, and Humphrey (2000) has compared the social functions of dreaming to those of play. The possibility that dreaming simulates pro-social and ag-

² For a recent exchange, see Hobson & Schredl (2011) and related commentaries in the International Journal of Dream Research (2011, vol. 4).

gressive social interactions in distinct sleep stages, and that these simulations might exert a regulatory influence on our waking social lives, was put forward by [McNamara et al. \(2005\)](#). Last, [Franklin & Zyphur \(2005\)](#) have considered how the simulation function of dreams might be expanded to cover social cognition and complex socio-cultural situations.³

The problem with the above social simulation theories of dreaming is that either they are not detailed enough to be testable, or that few, if any, have ever been directly tested against competing theories. They are interesting general ideas, but not strictly formulated theories that could be directly tested, or from which detailed predictions and potential explanations for the social contents of dreaming could be derived. Thus, these theoretical ideas have not led to a strong empirical, hypothesis-driven research program that would be able to systematically test the plausibility of these theories.

Whenever we formulate theories of dreaming, or of the functions of dreaming, they should be formulated in such detail that *empirically testable predictions* can be derived from them. Statements that are too vague or too general (e.g., “dreams are continuous with waking life”; “dreams are social simulations”) are difficult to test as such. The predictions derived from general statements are too unspecific. Thus, the theories remain uninformative but of course consistent with almost anything we might realistically expect to find in dream content. If a theory makes no detailed, risky predictions about what should or should not be found in dream content (under some specific circumstances or in specific populations) it doesn’t have much explanatory power, either. So far there is no detailed, convincing, testable theory

of the nature and the function(s) of social simulations during dreaming. There is also a lack of data on the detailed quantity and quality of simulated social interactions in dreams, and how they relate to real social interactions in the waking life of the same person. In the rest of this paper, we will try to outline ideas for the theoretical basis of a social simulation theory of dreaming and to formulate some empirically-testable hypotheses directly derived from the theory.

6 Towards a testable social simulation theory of dreaming

The relatively loose idea or the general observation that dreams are social simulations needs to be turned into a theory from which testable predictions can be derived. There are several ways in which this could be done. In the rest of this paper, we will formulate some suggestions towards that end. The basic assumptions that we adopt are based on the earlier work on the definition of dreaming (and consciousness) as an internal world-simulation in general ([Revonsuo 2006](#)). Any plausible theory of social simulation should also take into consideration, and draw from, concepts and advances in the fields of social psychology and evolutionary biology, in order to create a credible theoretical context into which social simulations in dreams can be placed. We will therefore connect the idea that dreaming may function as a platform for simulating social perception and interactions to some influential evolutionary biological and social psychological theories, as well as to the earlier simulation theory of the original evolutionary function of dreaming, the TST ([Revonsuo 2000](#)).

The two generally-accepted theories in evolutionary biology that seem to be relevant for the formulation of an evolutionary SST of dreaming are the Inclusive Fitness and Kin Selection Theory ([Hamilton 1964](#)) and Reciprocal Altruism Theory ([Trivers 1971](#)). Both are general evolutionary biological theories that apply not only to humans, but to multiple other species as well. Further, both have received ample empirical support from animal and human stud-

³ Another popular theory of dreaming postulates that the realistic simulation of character-self interactions serves the function of *emotion regulation* during dreaming ([Nielsen & Lara-Carrasco 2007](#)). In this group of theories, the function of dreaming is proposed to be the calming down of emotional surges, such as we see in psychotherapy ([Hartmann 1995, 1996, 1998](#)), or as reflecting the extinguishing of fear memories ([Nielsen & Levin 2007](#)). It is increasingly apparent that sleep plays a role in the consolidation of emotional memories, but whether sleep also *regulates* the emotional charge and valence of memories is not yet entirely clear (for a recent review, see [Deliens et al. 2014](#)). Thus, whether the emotional regulation theory has specific implications or predictions for social simulations in dreaming is not evident.

ies, and could thus serve as solid ground in guiding our thinking about social behaviours in evolutionary biological terms.

The Inclusive Fitness Theory ([Hamilton 1964](#)) postulates that an individual's genetic reproductive success is the sum of that individual's direct reproduction and the reproduction of the individuals carrying identical gene alleles. An individual can improve its overall genetic success by engaging in altruistic social behaviour that is directed towards individuals carrying identical alleles. The Kin Selection Theory is a more specific form of the inclusive fitness theory, which requires that the shared alleles are identical by descent. Thus, Kin Selection Theory postulates that an individual can increase its inclusive fitness by directing acts of altruism specifically towards genetic relatives, whereas inclusive fitness as such is not limited only to cases where kin are involved. Both, however, predict that acts of altruism should more often be directed towards individuals who share identical alleles.

Reciprocal Altruism ([Trivers 1971](#)) is defined as behaviour whereby an individual acts in such a way that temporarily reduces its fitness while increasing another individual's fitness. However, individuals engage in altruistic behaviour with the expectation that the recipient of the altruistic act will act in a similar manner at a later time. A strategy of mutual cooperation may be favoured when there are repeated encounters between the same individuals. Although cheating might be more beneficial for the individual in terms of immediate rewards, co-operation might provide net gain compared to short-term benefits.

Since selection pressures act on the typical conditions present in the history of any species, consideration of the demographics of the typical evolutionary environment of humans is crucial for understanding the evolution of social behaviours in our species. Recently, [Hill et al. \(2011\)](#) analyzed co-residence patterns among thirty-two present-day foraging societies, assuming that these might reflect an ancestral human group structure. They found that primary and distant kin of an adult individual accounted for approximately 25% of the co-resident adult

members of a band, i.e., about 25% of adult members in the group were directly genetically related, whereas about half of the adults were related through spouse or siblings' spouses, and the other 25% of adults were genetically unrelated.

If we accept the assumption that this observed distribution of relatedness approximates the degree of relatedness in ancestral human bands, there have been ample opportunities for ancestral humans to be subjected to selection pressures that could be explained using strategies postulated by the inclusive fitness and Kin Selection Theory, as well as Reciprocal Altruism Theory. There is ample evidence that people are more likely to help their relatives than genetically unrelated individuals (e.g., [Burnstein et al. 1994](#)), and that lethal violence is more frequently directed towards genetically-unrelated individuals than relatives ([Daly & Wilson 1988](#)). People also tend to be more altruistic towards other people in single round prisoner's dilemma game than could be expected ([Frank et al. 1993](#)) in order to protect their reputations. This seems to be a reasonable course of action, given that the faces of individuals labelled as untrustworthy cheaters are better recalled than those labelled as cooperative ([Mealey et al. 1996](#)). There are also rather large interindividual differences in altruistic behaviour, depending on factors such as age, sex, tendency to empathize, and circumstantial conditions.

The social environment has afflicted strong selection pressures on human cognitive faculties, and there are several theories that consider our essentially social nature. [Dunbar \(1992, 2008\)](#) has forwarded the Social Brain Hypothesis, which states that the main factor in the increase of our neocortical volume has been the cognitive demand bestowed on us by the increase in hominid group size. [Sutcliffe et al. \(2012\)](#) propose the idea that the costs and benefits of social interactions have been a critical driver for cognitive evolution. While our most intimate relationships are a source of social support, they are also the most costly as the quality of these relationships is dependent on the time invested in creating and maintaining them

over time. Forming weaker and less time-consuming ties with acquaintances can provide benefits such as information exchange and access to resources without exhausting an individual's resources that are allocated for social interaction. Our individual social worlds thus consist of hierarchically-layered sets of relationships defined by relationship intimacy, and different relationship types are designed to have different kinds of functions.

Turning our attention to the potentially relevant literature in social psychology, some further concepts and measures might be considered useful for dream theory. When it comes to the simulation of social interaction, one of the most relevant concepts is the social "Need to Belong" (Baumeister & Leary 1995). This fundamental motive towards interpersonal attachment and close, supportive social bonds pervades and influences our actions, emotions, and cognitions, and is fulfilled only by social affiliation and acceptance. To help us navigate the complex social world, and attune us to socially relevant information, two further advancements have been hypothesized in the form of the Sociometer Theory (Leary et al. 1995) and the social monitoring system (Gardner et al. 2000). Sociometer Theory proposes an internal monitoring device that feeds forward information about our level of social inclusion in the form of self-esteem or self-worth (Leary et al. 1998), whereas the social monitoring system is purported to guide the processing of social information whenever people's needs to belong are not being met (Pickett et al. 2004). In sum, the concept of "Need to Belong" in general, and the suggested social monitoring systems in particular, might prove useful in postulating testable hypotheses for the functions of social simulation in dreams. The Sociometer, for example, might act in a similar fashion to the threat cues postulated in TST, and prompt dreams to simulate relevant social skills or interactions.

An interesting developmental suggestion about the interplay between simulation mechanisms and social deficits has recently been put forward by Oberman & Ramachandran (2007), who propose that in typically developing individuals the abilities of Theory-of-Mind

(ToM), empathy, perceptual recognition, and motor mimicry might be mediated by an internal simulation mechanism or mechanisms. By taking into consideration a condition—autism—where all these abilities appear to be impaired, they make the case for a possible link between deficient simulation mechanisms and behavioural and social deficits. The exact implications of this idea for the hypothesis that dreams serve a social simulation function requires further consideration. One possibility is to test whether individuals with Autism Spectrum Disorders (ASD) dream less of social interactions, or whether their dreams of social interactions are different in content from those of other people. Thus far this line of research has not been explored in depth. Daoust et al. (2008) have looked into the dream contents of people with ASD, and found that they report significantly less dream-characters and social interactions than the control group. They note, however, possible error sources in the testing procedure, such as, for example, how the reporting of dreams itself might be affected by ASD.

There has been some research linking the effects of attachment relationships to dreaming. If, as attachment theory proposes, we use our early experiences with primary caregivers and other attachment figures as model states for future social interactions and the way we view and attune to our social world, it could be assumed that this would also affect our simulations of this world. Early attachment and bonding are, after all, quintessential for our species, and according to Fonagy & Target (1997) might also work as the basis for our abilities to mentalize or to create a ToM. McNamara (1996) has developed the idea that REM sleep is the mechanism that activates and maintains early attachment relations, as well as pair-bonding in later life. Selterman & Drigotas (2009) have found that attachment style is correlated to dream emotions when dreaming about romantic partners, so that those with anxious or avoidant attachment styles reported more stress, conflict, and negative emotions.

In an exploratory study on the dream contents of those suffering from Complicated Grief

(CG) after the loss of an attachment figure, [Germain et al. \(2013\)](#) found the dreams containing family members to become significantly more frequent, while there was no marked increase in the occurrence of deceased characters. Males suffering from CG also reported more familiar persons in their dreams than the control group. Both male and female CG patients also exhibited fewer negative emotions and fewer instances of aggression in their dreams, and females also had decreased amounts of positive emotions and friendliness.

We can thus conclude that the inherently social nature of our species is deeply ingrained, and has likely been as important for our survival in the ancestral environment as threat perception and avoidance skills. SST can therefore be formulated in an analogous manner to TST, but in addition to the evolutionary background theory, also taking into consideration important social functions such as the need to belong, social bonding, social networking, and social support as essential ingredients.

TST ([Revonsuo 2000](#)) places the contents and the function of dreaming in an evolutionary-psychological context and proposes that dreams were selected for their ability and propensity to simulate threatening events in a safe way, thus preparing the individual to survive real-life dangers. The hypotheses and predictions of the TST, especially concerning the inclusion of threat simulations in dream content, have gained support from several independent sources, such as studies on the content of nightmares and bad dreams (e.g., [Robert & Zadra 2014](#)), recurrent dreams ([Valli & Revonsuo 2006](#); [Zadra et al. 2006](#)), post-traumatic dreams in children and adults ([Bulkeley & Kahn 2008](#); [Valli et al. 2006](#)), dreams anticipating a stressful experience ([Arnulf et al. 2014](#)), children's earliest dreams ([Bulkeley et al. 2005](#)), dreams and mental contents in parasomnias ([Ugucioni et al. 2013](#)), the dreams and nightmares of new mothers (which mostly depict the infant in peril and trigger protective behaviours, [Lara-Carrasco et al. 2013, 2014](#); [Nielsen & Lara-Carrasco 2007](#)), as well as dreams of the general population (for a review, [Valli & Revonsuo 2009](#)).

Thus, when it comes to emotionally negatively-charged dream contents that simulate some sort of dangerous situation or unfortunate event, the TST seems able to quite well predict and explain many features of the quantity and the quality of the threat simulations found in the data. Therefore, a similar theoretical approach might also prove fruitful in the case of social simulation theory. The SST, however, needs to be formulated in such a manner that its predictions can be clearly distinguished from those of the TST.

As negative and threatening events commonly occur in dreams, the TST alone already covers a fairly large proportion of dream content. But it also ignores a relatively large proportion of dream content, as it does not offer any explanation of non-threatening dreams or for the simulation of neutral and positive events in dreams. This raises the question: do types of dream events other than those that are threatening have some evolutionarily-based simulation function, independent of the threat-simulation function of dreaming? Are there events that are equally important targets for simulation as the negative, threatening situations simulated in threat simulation dreams?

TST covers threatening events in dreams, whether social in nature or not. Many threatening events of course do involve social interaction (such as verbal or physical aggression), but are explained by the TST as primarily simulations of specific types of threat, and therefore as rehearsals of threat perception and threat-avoidance behaviours, rather than as simulations of social interactions as such. A social simulation theory that explains dreams that TST does not cover should thus focus on social simulations that are largely independent of the threat-simulation function. In some dreams these two types of simulation may, however, be difficult to tease apart. For example, a social simulation theory might account for some social interactions that happen during a threatening event in a dream, such as how the Dream Self interacts with others and collaborates with them during a threatening situation. Furthermore, these two simulation theories may not be mutually exclusive but instead complement each other. Some specific

types of simulations of negative social interactions are better accounted for by the TST while other, positively toned simulations can be explained by the SST. For example, from an evolutionary perspective it might make sense to simulate different kinds of interactions, friendly or aggressive, with people belonging to different layers of our social hierarchy.

We are open to the possibility that social simulation is an original evolutionary function of dreams alongside the threat-simulation function of dreaming. We believe that social simulation theories hold much promise. But before this belief can be empirically justified, a testable version of the social simulation theory needs to be formulated. Such a theory should independently cover the social simulations in dreams that fall outside the scope of the TST.

Furthermore, also the predictions of the CH must be distinguished and separated from those of the SST. Therefore, the question becomes: What aspects of human social reality might dreams be specialized in simulating in such a way that these social simulations have significant consequences for cognition and behaviour during the waking state, and in virtue of which social simulations during dreaming have fulfilled important functions in the evolutionary history of the human species? What kind of social-cognitive processes and behavioural social skills might have been both critical enough both for an individual's survival and successful reproduction, as well as occurring frequently and universally enough in the human ancestral environment, to be selected for as a universal feature of human dreaming? Moreover, those processes and skills would have to be something that in fact *can* be regularly simulated by the dreaming brain, and they have to be contents that actually *are* being simulated frequently and universally in human dreaming, according to the evidence from content analysis studies of dreaming.

To sum up, a credible version of the SST should have predictions and explanations that are clearly different from both the TST and the CH. To be different from TST, the SST should predict and explain the social simulations that happen outside threatening events in dreams, and to be different from the CH, the SST

should predict that some types of social stimuli, social cognition, or social behaviours are simulated actively and selectively, so that they are overrepresented in dreams as compared to waking life.

We will first consider some basic cognitive processes that might fulfil these roles and will then proceed to more complex social behaviours and interactions. We admit that many of these ideas are at this stage speculative. But if it is possible to formulate them in an empirically testable manner, then we can figure out later on which ideas remain mere empirically unsupported speculations, and which ones might actually predict and explain central aspects of our dream content.

6.1 The simulation of social perception as a function of dreaming

Overall, there are good reasons to support the view that fast and errorless social perception abilities were universally important skills for humans during their evolutionary history, and, therefore, rehearsing them through dream simulations would have served to maintain and enhance their speed and accuracy during wakefulness. In the ancestral environment, fast and efficient social perception and recognition mechanisms were essential for telling friends and allies apart from potential enemies. Thus, detecting the presence of other human beings in the same spatiotemporal context where oneself is located, immediately classifying them in terms of familiarity, identity, and history of past interactions with them, and predicting the nature of future encounters with them must have been an important survival skill. Perhaps it was important enough that rehearsal of these social-cognitive functions through social simulations during dreaming would have increased an individual's inclusive fitness.

The social perception system needs to quickly estimate answers to the following questions: *am I alone in here or are there other humans present? Are the other humans around me familiar to me or are they strangers?* Thus, the first stage of social perception is to detect other humans in the vicinity and to classify them in

terms of unfamiliar people (strangers) vs. familiar people. As [Diamond \(2012\)](#) explains in “The World Until Yesterday”, in most traditional societies during human evolutionary history, to encounter strangers was unusual and typically considered potentially dangerous, because the social interaction that followed might not necessarily have been peaceful in nature.

The second stage of social perception deals in more detail with the familiar people that are detected. If the people in my presence are familiar to me, who exactly are they? What is my relationship with them? What have my past interactions with them been like? What should I expect the interaction between us to be like this time around? To answer these questions, familiar people need to be quickly identified. Based on semantic and autobiographical memory information that we have about people familiar to us, we quickly activate expectations and strategies as to how we should interact with the people around us in the most constructive way.

But so far this idea is mere speculation. What kind of *testable hypotheses and predictions* could be derived from this theory? How could we derive predictions that clearly distinguish the SST from the CH? The CH does not attribute any evolutionary simulation functions to dream content; according to CH, dreaming simply and passively *mirrors* whatever experiences have recently been encountered in the dreamer’s waking life (and thus impressed on long-term memory). Obviously, therefore, it would not lend sufficient (or specific) support to the SST to predict that social perception should be found in dreams in the same proportions as in waking life, because the CH predicts and explains exactly the same observation and, moreover, does it more parsimoniously, without postulating any just-so-story of evolutionary *functions* to social dream content.

The SST must thus go beyond the CH and make the risky prediction that, if social perception is the original evolutionary function of dreaming and it is therefore still expressed in our dream contents, then dreams are *specialized* in simulating social perception. If dreams are specialized in simulating social perception, then perceptual contents, cognitive processes, and

behaviours relating to social perception skills should occur (as simulations) in a selective or *exaggerated* form in our dreams. The testable prediction derived from this is that during dreaming, social perception occurs *more frequently* than in waking life (shows quantitatively an increased frequency) and/or qualitatively in a more difficult or challenging form than in waking life.

Quantitatively, dream simulations could exaggerate the proportion of the types of stimuli that were most important to recognize quickly and accurately during evolutionary history (e.g., strangers vs. familiar people; enemies vs. friends). It is important to process this information quickly because the information had high survival value in ancestral environments. Furthermore, dream simulations could present qualitatively challenging stimuli for the social perception system; for example, more variety of different kinds of stimuli (different kinds of familiar and unfamiliar simulated people), or ambiguous stimuli that are more difficult to perceive or interpret than real life stimuli (vague or unstable simulations of people).

Conversely, if the social stimuli in dreams simply mirror the social stimuli during wakefulness (and memory representations of them), quantitatively and qualitatively, then the CH gains support: dream experiences merely *copy* the patterns and rates of social stimulation encountered during wakefulness, but do not *selectively* and *actively* simulate them in ways and proportions that would reflect some original evolutionary functions and would therefore have supported important survival skills in ancestral environments.

To test these two opposing theories, SST and CH, against each other empirically, we need detailed information not only about the quantity and quality of social perception in dreams, but also about the quantity and quality of social perception during wakefulness in the same subjects’ lives during the same period of their lives. Some studies already exist that provide us with this kind of data, but most of the hypotheses remain to be tested in future studies that should be explicitly designed to test the opposing hypotheses and predictions of the two theories.

McNamara et al. (2005) conducted an interesting study that can be interpreted as testing the SST prediction that social perception is quantitatively exaggerated in dreams as compared to waking life. They conducted experience sampling from fifteen individuals over two weeks across waking, REM sleep, and Non-Rapid Eye-Movement (NREM) sleep states. The participants recorded verbal reports of their perceptual and other experiences when paged at random intervals during sleep or wakefulness.

The results showed that *more characters appeared in dreams than in wake reports*. Unfortunately McNamara et al. (2005) do not report the exact descriptive statistics of this finding, so we do not know how large this difference exactly was. In any case, this finding is better in accordance with the predictions of the SST than CH: Stimuli requiring social perception (human characters) are present at higher frequencies during dreaming than during wakefulness, when experiences from both states are sampled and reported in a similar manner.

This important finding suggests that the basic processes and skills required in social perception are more engaged during dreaming than during an equal stretch of time in wakefulness. This lends support to the hypothesis that *dreaming is specialized in the simulation and rehearsal of social perception*, which may thus be one of the original evolutionary functions of dreaming. It has to be added, however, that McNamara et al. (2005) is the only study so far that provides us with this kind of data, where the frequencies of the social contents of dreaming and waking experiences have been directly compared with each other. Replications are obviously required in different populations and in larger samples of dreams and waking experiences. But so far, so good for SST.

The same study can be taken to test the additional prediction of SST, namely that dream simulations of human characters should exaggerate the proportion of the particular types of stimuli that were, during evolutionary history, most important to recognize quickly. Meeting strangers posed a threat in the original evolutionary context; thus, the SST predicts

that *strangers or unfamiliar people should be overrepresented in dreams as compared to waking life*, to simulate and rehearse the type of perceptual categorization (familiar vs. unfamiliar) that was most important in the evolutionary context. McNamara et al. (2005) report that the proportion of strangers (or unfamiliar people) encountered in dreams is indeed significantly higher than in waking life. Only 25% of people present in the waking episodes were unfamiliar, whereas about 50% of the (simulated) people in dreams were unfamiliar. Again, this discrepant pattern is well predicted by and accounted for by the SST, but goes against the predictions of the CH.

The recognition and identification of familiar people as who exactly they are could also potentially be a target of useful simulation in dreams. It might be argued from SST that quick and correct recognition of familiar people enhances the quick selection of the appropriate social strategies and behaviours when we interact with them. As about 50% of simulated people in dreams are familiar, there are still plenty of opportunities to rehearse these recognition skills. There are, however, no studies that would have directly and quantitatively compared the frequency of face recognition during dreaming and wakefulness. But still, there are some studies that question whether face recognition is engaged during dreaming and to what extent.

Kahn et al. (2002) report, in a character recognition study, that about 45% of familiar dream characters were recognized through their appearance (including facial features), and an additional 12% by their observable behaviour. Thus, nearly 60% of dream characters are recognized perceptually. However, about another 12% of dream characters are recognized intuitively, by “just knowing” who they are, which suggests that in those cases, the “recognition” happens in a top-down manner and is therefore independent of the perceptual and facial features of the dream character.

If familiar persons are *not* overrepresented in dreams to begin with (as the McNamara et al. 2005 study suggests), and only well *under 50% of the familiar people simulated in dreams*

are recognized through their facial features, this pattern of data does *not* particularly support the idea that dreams are specialized in rehearsing familiar face recognition. However, we still lack knowledge about the frequency of face recognition in waking vs. dreaming, and only a study directly making that comparison could properly test this idea. So, the case remains open, but the expectations are not particularly high that this prediction of the SST will gain strong support in the future.

6.2 The simulation of mindreading as a function of dreaming

In addition to the processing of familiarity and identity, another aspect of social perception is called Theory-of-Mind (ToM) or “mindreading”. This refers to the interpretations we automatically make about the internal mental states of the people around us. We not only categorize the people around us as familiar and unfamiliar, and assign an identity to familiar persons, we also attribute thoughts, beliefs, motives, and emotions to them. As mindreading is crucial for our ability to predict and explain other people’s behaviours, our mindreading abilities could potentially have been a target of simulation during simulated social perception in dreams (Kahn & Hobson 2005; McNamara et al. 2007).

The study by Kahn & Hobson (2005) quantifies the frequency of mindreading activities in dreams. In one sample of thirty-five participants and about nine dream reports per participant, about four dream characters per report were observed on average. In over 80% of these dreams, the participants reported having had engaged in mindreading (at least one of) the other dream characters’ internal mental states. In another sample, 24 subjects reported on average six dreams per participant. Each dream was divided into separate dream events (on average four events per report were found), and the participants were asked to report, concerning each event, whether or not they were engaged in mindreading the other dream characters. In 50% of the episodes, mindreading was reported to have occurred. Thus, on the basis of these results, we may say that mindreading fre-

quently occurs during dreaming. Kahn & Hobson (2005) in fact suggest that this may be evidence for a specific simulation function being at work:

The two studies undertaken here support the idea that dreaming may provide a simulation of waking life as suggested by Revonsuo (2000), though not restricted to only threatening events. Instead, the data of these studies suggest that if dreaming is a simulation process, it is a simulation that provides a way of knowing and dealing with the intentions of others, both positive and negative. (p. 56)

The above studies show that mindreading is well represented in dreams, but they cannot tell us whether mindreading is *overrepresented* in dreams, as its frequency of occurrence cannot be directly compared to waking life. However, McNamara et al. (2007) have conducted a direct comparison of the frequency of mindreading between waking experiences, REM dreams, and NREM dreams of the same subjects. This is what they found:

REM reports were three times as likely to contain instances of mind-reading as were wake reports and 1.3 times as likely as NREM reports. Of 100 reports per state, there were 39 instances of mind-reading in REM reports, 29 in NREM reports, and 12 in wake reports. (McNamara et al. 2007, p. 211)

In conclusion, from looking at these studies, we may say that mindreading activities frequently occur in dreams, and that their frequency of occurrence is significantly greater during dreaming than during wakefulness: Mindreading is overrepresented or exaggerated during dreaming. Thus, this data *supports the SST prediction that dreaming specifically simulates mindreading* in order to maintain and rehearse our mindreading abilities, rather than the CH prediction that dreaming simply reflects the amount of mindreading we engage in during wake experiences.

Another finding that might indirectly lend support to the SST-mindreading idea is that the behaviours and communications of dream characters are often bizarre (Kahn et al. 2002; Revonsuo & Salmivalli 1995; Revonsuo & Tarkko 2002); that is, they are unusual, unexpected, and thus unpredictable on the basis of our waking expectations. Studies on intentional social interactions between the Dream Self and other avatars in lucid dreaming suggest that dream characters are largely independent of the dreamer and behave autonomously (Stumbrys et al. 2011; Tholey 1989). Unusual and unpredictable behaviours could be interpreted simply as failures of the dream simulation to produce credible sequences of real-life behaviour. But they could also be interpreted as particularly engaging and activating social stimuli that serve to challenge our mindreading skills. That is, bizarreness in this case could be functional in the sense that it makes the simulation more challenging. Perception of unexpected behaviours may trigger a reconsideration of what is going on in the character's mind in order to produce such unexpected behaviour, and thus present a frequent need to engage in mindreading as we interact with unpredictable characters in our dreams. This idea could be empirically tested by studying whether bizarre behaviours on the part of dream characters tend to trigger mindreading in the Dream Self, and whether this feature of dreams might partially explain the apparently frequent engagement in mindreading in dreams.

6.3 The simulation of social interactions as a function of dreaming

Humans are an essentially social species and an individual's survival in the ancestral environment was most likely entirely dependent on the individual's ability to form long-lasting positive social bonds with close kin and other group members who offered protection, access to nutrition and other crucial resources for survival, collaboration, friendship, social support, mating opportunities, and opportunities to gain a better social status within the group.

Social interaction in dreams is a more complex affair than simple social perception. There need to be some behaviours that link dream characters and the Dream Self, where the intentional behaviour of one character (or the Dream Self) is directed at another character (or at the Dream Self), and the recipient somehow registers it or reacts to it. Traditionally, in the Hall & Van de Castle (1966) content analysis system, social interactions have been classified into three different categories: aggression, friendliness, and sexual interactions. It may be, however, that these three categories are too broad, and do not cover or identify all theoretically-interesting types of social interaction.

When it comes to the simulation of social interactions, the predictions of the SST should, again, be contrasted with the predictions derived from competing theories. In this case the SST needs to be distinguished from two other theories: CH and the TST. The TST is a simulation theory that describes and explains the simulation of aggressive behaviours in dreams, by including them under the category of "threatening events". The function of dreaming, according to TST, is not to specialize in the simulation of social interactions *per se*, but in threatening events; thus, any social interactions are simulated in dreams not because they are social events but because they are threatening events. No independent social simulation theory is required to explain the simulation of social interactions involving a threat; and aggressive behaviours between dream characters are, obviously, social interactions where the wellbeing of the Dream Self or some other dream character is potentially threatened.

Compared to CH or SST, the TST can account for the overrepresentation of threatening events and aggressive interactions in dreams (as compared to waking life, McNamara et al. 2005; Valli et al. 2008). The TST, however, gives no description or functional explanation for neutral and positive types of social interactions (unless they occur as parts of a threatening event). The TST assumes that neutral and positive events in dreams are either parts of a threat simulation (e.g., responding to a threat by helping others who are targets of a threat) or that they repres-

ent some kind of superfluous, non-functional dreaming that simply goes on automatically even if the threat simulation mechanisms are not activated. Thus, when it comes to social interactions, the SST should in particular predict and explain the neutral and friendly types of social interactions, and show that some of them are actively selected as targets of dream simulation. In contrast, the CH predicts that neutral and positive types of social interactions should only occur in the same proportions as they occur in real life, passively reflecting their waking-life frequencies.

If, according to SST, the simulation of neutral and positive social interactions in dreams serve to represent and strengthen important social connections and to rehearse prosocial behaviours in relation to those connections, then these types of interactions should frequently occur in dreams. This would serve the function of maintaining, rehearsing, or strengthening our waking life social bonds and networks, and would satisfy our social need to belong to groups that enhance our survival. After dreaming about prosocial behaviours, our social bonds during wakefulness would automatically be experienced as stronger and we would be more likely to engage in behaviours that further strengthen those bonds. Some tentative steps towards examining how the affects and contents of social dreams predict subsequent waking behaviour have been taken by [Selterman et al. \(2014\)](#). They discovered that an increased frequency of dreams involving significant others was associated with higher levels of intimacy and interaction the following day, whereas dream infidelity predicted less intimacy. Reported arguments in dreams were also found to be correlated with subsequent conflict in waking life. They leave open the question whether this is due to the conscious reflection of the reporting procedure, a more implicit association, or a mixture of the two.

Again, there are no detailed content analysis studies that have investigated the exact nature of social interaction in dreams by taking into account the social context of the interaction; that is, by studying who is engaged in what type of interaction and with whom. From

previous studies based on home dream diaries we know that dreamer-involved aggression, adjusted to take into account all social interactions except sexual interactions, is present in 60% of male dreams and half (51%) of female dreams ([Domhoff 1996](#)). When male strangers appear in a dream, the likelihood that physical aggression will occur in that dream far exceeds what would be expected on the basis of chance. Basically this means that male strangers signal physical aggression. The dreamer, however, is an aggressor in 40% of male dreams and a third of all female dreams ([Domhoff 1996](#)).

Yet, as the Hall and Van de Castle norms indicate, there are friendly interactions in dreams—slightly more often in female (42%) than male (38%) dreams ([Domhoff 1996](#)). Females also dream more often of familiar people (58%) than of strangers (42%) while the opposite is true for males (45% vs 55%, respectively); which might suggest that when there are more familiar people in dreams, there is also more friendliness. The dreamer participates in the majority of interactions that involve friendliness (84% for females, 90% for males), and the befriender proportion is 50% for males and 47% for females. Thus, both sexes initiate friendly interactions in their dreams approximately as often as they are befriended. Helping and protecting is the most frequent type of friendly behaviour in both sexes, followed by friendly remarks and compliments, and giving gifts or granting loans. Surprisingly, however, there is very little mutual or reciprocal friendliness, so although friendly interactions are initiated in dreams by the Dream Self or other characters, in less than 10% of friendly interactions the act is reciprocated immediately. This observation goes against any social simulation theory that predicts reciprocal friendliness should be highly represented in dreams: this does not seem to be the case.

[McNamara et al. \(2005\)](#) investigated whether types of social interaction are different in REM than in NREM dreams compared to wakefulness, and noticed that aggressive interactions were more often simulated in REM dreams, whereas friendly interactions were more often simulated in NREM dreams. Furthermore,

dreamer initiated friendliness was more typical for NREM than REM dreams. What is most interesting in this study, however, is that they also found that social interactions in general are more often depicted in both REM and NREM dreams than in wake reports. While aggression was more often simulated in dreams than encountered in waking life, the number of reports with at least one occurrence of friendliness did not differ significantly across sleep–wake states. Thus, these observations imply that dreams do not seem to overrepresent friendly interactions as compared to waking experiences.

In sum, aggressive interactions seem to be more prominent in dreams than neutral or friendly interactions, which would lend more support to the TST than to SST, and friendly interactions are not more prominent in dreams than in waking life, which would lend support to CH and the TST. Nevertheless, if simulations are biologically functional, and if these two types of simulation functions are not mutually exclusive, might there be enough room in the dream content for simulation of neutral and positive interactions, in such a way that it could have contributed to the inclusive fitness of our dreaming ancestors?

6.4 Some testable ideas derived from SST

Let us see how this general approach to social simulation in dreams could be translated into some directly testable hypotheses. Now, a general thesis derived from the SST could be formulated as follows:

Dreams are *specialized* in simulating *the most important social connections and networks* of the dreamer to give an additional selective advantage and to enhance the survival of the dreamer in waking life. The simulations of particular people (the frequency of their presence in a person's dream life), and the simulations of positive interactions with particular people, should focus on the people closest to us in waking life and on the social bonds most important for our inclusive fitness in the real world.

This thesis could be directly tested by deriving some empirical predictions from it, telling

us what kind of simulations of social interactions and to what extent they should appear in dreams. If dreams are specialized in the way predicted by SST, then the most important social networks and the people in them *should appear more frequently in dream life than in a corresponding stretch of waking life*. That is, their frequency of occurrence should be targets of active selection and inclusion into dreams, and hence over-represented and exaggerated in dreams.

This empirical prediction could be tested by identifying a person's most important social networks in waking life, and by quantifying the frequency of interactions of the dreamer with those people during dreaming vs. during wakefulness. In the already existing literature, there are some data relevant to the hypothesis, but data that directly compares waking social life and dream life in the manner required to test the hypothesis seems to be lacking.

The data scattered in the literature describes the relative frequency of dreams in which a certain type of close person appears on average in the dreams of the general (or the student) population. For example, romantic partners occur in 20% of dreams and this frequency correlates with the time spent together in wakefulness (Schredl 2011; Schredl & Hofmann 2003). Core family members occur in 10%–30% of dreams; parents in about 8%–20% of dreams, and siblings from 2%–7.5% of dreams (see Schredl 2013). Friends occur in about 20% of dreams (Roll & Millen 1979), but during long-term isolation from social contacts with friends in one case (Merei 1994) this declined to 10%. In studies of long dream series from a single person, a close family member or spouse has been found to be the person most often dreamed about. In a sample of over two hundred dream reports, reported by a married woman (Arlie) with four grown-up children, the most frequently occurring character is her husband; whereas in a sample of over three hundred dreams from an unmarried woman in her thirties (Merri), the most frequently occurring character is her sister, who was no longer alive at the time when the dream reports were collected (Schweickert 2007).

In Schredl's studies, interesting analyses of a long dream series from a single dreamer were conducted, revealing the proportions of schoolmates (2012) and family members (2013) simulated in dreams across a period stretching over twenty years. Old school mates continued to appear in about 5% of dreams over the years when the dreamer had nothing to do with them any more in real life. Similarly, family members, even when the participant was not living with them anymore, still retained a strong if somewhat reduced presence in the same dream series, being present in approximately 15–20% of the dreams over a twenty-year period.

These results show that the probability of occurrence of a character in dreams is to some extent related to the amount of real life contact with that person and to the closeness of the relationship in real life, thus supporting the CH. However, people who have at some point in life been close and important *do not seem to disappear totally* from the dream simulations even though they have long ago totally disappeared from the real life of the dreamer. This feature of the already-existing data suggests that simulations of social contact might serve the function of maintaining or strengthening close relationships over time. When the frequency of a previously close and important social contact falls to zero in waking life, and the person is no longer encountered in waking life (like old school mates after leaving school, or after the death of a family member), the simulation of such a person seems never to totally disappear from dream life, even if the frequency of dream simulations of that person to some extent diminishes. Social simulations in dreams thus seem to maintain an active storage and rehearsal of the most important and closest social relationships of our entire lives, even when those relationships are broken or discontinued for good, or are temporarily on hold in our waking lives.

What happens if a relationship that has disappeared from waking life is reactivated after years of disconnection? In Schredl's (2012) study, old schoolmates met for a reunion twenty years after going their separate ways. Interestingly, when the same relationships are re-activated in real life for just one day, the dream sim-

ulation of those social relationships is increased significantly and for a long period of time (compared to the time of actually meeting). The mechanism that reactivates old targets of simulation might be analogous to that proposed in TST for the re-activation of old threats. The frequency with which the most important real threats are simulated (e.g., in post-traumatic nightmares) increases when, during wakefulness, new cues are encountered that are associated with the old threat possibly reoccurring in real life.

These considerations suggest a more precise function of social dream simulations that could be formulated along the following lines. We may call it the Strengthening Hypothesis: *the function of social simulations in dreams is to maintain and strengthen the dreamer's most important social bonds from waking life*. Consequently, a prediction derived from the Strengthening Hypothesis can be formulated as follows: if strengthening important social bonds is a function of social dream simulations, then dreaming should include with high frequency social interactions in which the (current or past) most important social bonds are strengthened through various types of simulated positive social interactions and prosocial behaviours. Thus, the frequency of prosocial, positive interactions (bond-strengthening) with the most important persons should clearly surpass the frequency of negative (bond-weakening) interactions within dreams, and also be more frequent in dreams than in a corresponding stretch of waking life.

Schredl's (2012, 2013) findings are to some extent consistent with both the CH and the SST, but do not allow any firm conclusions about which theory better predicts the occurrence of the most important social connections in dreams. Studies that collect data from both waking life and dream life during the same period of life from the same people, as well as from the life history of these individuals, are necessarily required to test whether the representation of the most important connections is exaggerated in dreams, or if they just reflect the waking frequency. In practice, this prediction could be tested by identifying all the interactions between the dreamer and the people in his

or her most important social networks, in both dream and waking reports. Then the interactions could be classified according to whether they tend to strengthen or weaken the relationship with that particular person. If the frequency with which dreaming simulates positive interactions surpasses the frequency of those interactions in real life, then the SST would gain credence over the CH.

Another potential simulation function to consider can be called the Practise and Preparation Hypothesis. According to this hypothesis, the function of social simulations in dreams is to force the dreamer to *practise important social bonding skills*, such as how to give social support to others. The prediction derived from this hypothesis states that if practising social bonding skills is a function of dreaming, then the dreamer should frequently offer various types of social support to other dream characters, for example emotional, instrumental, or informational support. Furthermore, the types of social support offered should be dependent on the degree of relationship intimacy, i.e., the distance between the self and the recipient in the hierarchy of the social world of the individual. If the Practise and Preparation Hypothesis is correct, then the frequency of simulating social support should be higher than comparable behaviours in real life.

These ideas are testable, but dream content studies are to be carefully designed with the specific aim of testing them. In the literature already published, friendliness percentages in different dream samples and descriptive statistics concerning who initiates friendliness in dreams might shed some light on these questions. However, without any data about the frequency of occurrence of these same behaviours in the waking state of the same person, the purely descriptive findings from dreaming alone will not be able to separate CH predictions from SST predictions. The comparable waking data is crucial as a baseline against which the dream data can be evaluated and in relation to which the CH predictions can be contrasted with the SST predictions.

In an ideal setting the hypotheses for the SST and its proposed functions would also be

tested cross-culturally and in particular, as the theory makes bold evolutionary claims, in traditional small-scale human societies. As [Henrich et al. \(2010\)](#) have pointed out, the concentration of behavioural research into the so-called Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies are highly unrepresentative of the species, and might pose problems for the generalizability of the results. Furthermore, by contrasting, for example, the differences between the social simulations of small-scale and Western societies, we might uncover useful information about the plasticity and ontogenetic mechanisms of the social simulation function.

7 Conclusions

The concept of “simulation” is a useful theoretical concept for dream research. It unifies definitions and descriptions of the basic nature of dreaming, and helps to formulate testable theories of the function of dreaming. Applying this concept to the social reality of dreams means that we start to describe the persons and social interactions in dreams as simulations of their counterparts in real life. Consequently, we can ask: How does the simulated social reality relate to the actual social reality in the same person’s waking life? Is it plausible to hypothesize that the avatars in the dreaming brain might in fact be there in order to force us to maintain and practise various evolutionarily important functions of social perception and social bonding?

In this paper we made an attempt to clarify what it means to put forward the theoretical statement that “dreaming is a social simulation”, especially when this claim is offered as an expression of a theory of the *function* of dreaming. The SST can be formulated in a testable manner, and a number of testable predictions can be derived from it. Some of those predictions, concerning basic social perception and mindreading abilities, already receive rather strong support from the published literature. Many more hypotheses remain to be tested. To achieve theoretically-informative results and to directly contrast the predictions of different theories, future studies have to be designed in a

strictly theory-driven and hypothesis driven manner—which, unfortunately, is not a common approach in dream research.

If the SST, or some parts of it, prove successful, we have to be able to show that the SST predicts the nature and the occurrence of social simulations in dreams more accurately than its main competitors, the CH and the TST. To fare better than the CH, the data would have to show that the most important social contents are actively selected for incorporation in dreams as social simulations, and therefore rehearsed in an exaggerated quantity or form in dreams. To show that the CH is on the right track, the data would have to show that dream simulations merely reflect, both quantitatively and qualitatively, whatever experiences waking life has recently presented to the same person. To go beyond what the TST predicts and explains, the data supporting the SST would have to show that dreaming over-represents and actively runs positive or neutral social simulations in dreams that strengthen the skills of social perception and bonding, but that have nothing specifically to do with threat-perception and avoidance.

At this point, we are not yet sure how strong the empirical case for SST is going to be, and whether the evidence will mostly turn out to be for or against it. We shall wait for the kind of studies that directly test SST and set it against other theories' predictions. However, what we are confident about is that SST *is* an empirically testable theory, and that dream research would in general gain much if dream content studies were rigorously designed to test the predictions derived from opposing theories, and if dream data were in general collected and analysed in a manner that provides us with strong tests of different theoretical hypotheses rather than just producing more and more purely descriptive data of dream content (and then presenting vague, post-hoc theoretical interpretations of them). In that way, dream research would be able to find and test new, promising theoretical ideas, perhaps derived from cognitive and social neuroscience and from evolutionary psychological considerations. New theoretically-guided studies would help leave behind old

ideas if they did not generate any clear and testable predictions or if such predictions did not gain sufficient empirical support.

Even if we will at some point be able to explain some of the functions of social simulation in our dreams, we might not be able to explain the *underlying mechanisms that generate* the simulations. The fundamental metaphysical nature of the simulated persons inhabiting our dreaming brain might after all be almost equally mysterious as the immaterial nature of a Cartesian ghost, because, like everything we experience in our dreams, the avatars in our dreams are built out of features that have no objective, physically observable, or measurable substance. Instead, they consist of subjectively-experienced phenomenal features, and at least at the present state of consciousness science, the only way for us to get any empirically-based data about them is through the introspective reports carefully collected from the dreamers. How the sleeping brain produces vivid, dynamic, complex phenomenality and organizes it into subjective spatiotemporal hallucinations, inhabited by avatars and social simulations, still remains beyond any current theoretical explanations of dreaming and consciousness. Any plausible explanation of the actual brain mechanisms that do the trick would have to solve the hard problem of consciousness (Chalmers 1996) and cross the explanatory gap (Levine 1983) between the objective neural mechanisms in the brain and the subjective experiential realities going on in subjective consciousness. We are not quite there yet.

Acknowledgements

This research was supported by the Academy of Finland, Research program HUMAN MIND, project number 266434.

References

- Arnulf, I., Grosliere, L., Le Corvec, T., Golmard, J.-L., Lascois, O. & Duguet, A. (2014). Will students pass a competitive exam that they failed in their dreams? *Consciousness & Cognition*, 29, 36-47. [10.1016/j.concog.2014.06.010](https://doi.org/10.1016/j.concog.2014.06.010)
- Baumeister, R. F. & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, 117 (3), 497-529. [10.1037/0033-2909.117.3.497](https://doi.org/10.1037/0033-2909.117.3.497)
- Brereton, D. (2000). Dreaming, adaptation, and consciousness: The Social Mapping Hypothesis. *Ethos*, 28 (3), 379-409. [10.1525/eth.2000.28.3.379](https://doi.org/10.1525/eth.2000.28.3.379)
- Bulkeley, K., Broughton, B., Sanchez, A. & Stiller, J. (2005). Earliest remembered dreams. *Dreaming*, 15 (3), 205-222. [10.1037/1053-0797.15.3.205](https://doi.org/10.1037/1053-0797.15.3.205)
- Bulkeley, K. & Kahan, T. L. (2008). The impact of September 11 on dreaming. *Consciousness and Cognition*, 17 (4), 1248-1256. [10.1016/j.concog.2008.07.001](https://doi.org/10.1016/j.concog.2008.07.001)
- Burnstein, E., Crandall, C. & Kitayama, S. (1994). Some neo-Darwinian decision rules for altruism: Weighing cues for inclusive fitness as a function of the biological importance of the decision. *Journal of Personality and Social Psychology*, 67 (5), 773-789. [10.1037/0022-3514.67.5.773](https://doi.org/10.1037/0022-3514.67.5.773)
- Chalmers, D. J. (1996). *The conscious mind*. Oxford: Oxford University Press.
- Daly, M. & Wilson, M. (1988). *Homicide*. Hawthorne, NY: Aldine de Gruyter.
- Daoust, A. M., Lusignan, F. A., Braun, C. M., Mottron, L. & Godbout, R. (2008). Dream content analysis in persons with an autism-spectrum disorder. *Journal of Autism and Developmental Disorders*, 38 (4), 634-643. [10.1007/s10803-007-0431-z](https://doi.org/10.1007/s10803-007-0431-z)
- Deliens, G., Gilson, M. & Peigneux, P. (2014). Sleep and the processing of emotions. *Experimental Brain Research*, 232 (5), 1403-1414. [10.1007/s00221-014-3832-1](https://doi.org/10.1007/s00221-014-3832-1)
- Diamond, J. (2012). *The world until yesterday*. London, UK: Penguin.
- Domhoff, G. W. (1996). *Finding meaning in dreams: A quantitative approach*. New York, NY: Plenum.
- (2007). Realistic simulation and bizarreness in dream content: Past findings and suggestions for future research. In D. Barrett & P. McNamara (Eds.) *The New Science of Dreaming* (pp. 1-27). Westport, CT: Praeger.
- Dunbar, R. I. M. (1992). Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22 (6), 469-493. [10.1016/0047-2484\(92\)90081-J](https://doi.org/10.1016/0047-2484(92)90081-J)
- (2008). Why humans aren't just great apes. *Issues in Ethnology and Anthropology*, 3 (3), 15-33.
- Fonagy, P. & Target, M. (1997). Attachment and reflective function: Their role in self-organization. *Development and Psychopathology*, 9 (4), 679-700.
- Foulkes, D. (1985). *Dreaming: A cognitive-psychological analysis*. Hillsdale, NJ: Lawrence Erlbaum.
- Frank, R. H., Gilovich, T. & Regan, D. T. (1993). The evolution of one-shot cooperation: An experiment. *Ethology and Sociobiology*, 14 (4), 247-256. [10.1016/0162-2095\(93\)90020-I](https://doi.org/10.1016/0162-2095(93)90020-I)
- Franklin, M. S. & Zyphur, M. J. (2005). The role of dreams in the evolution of the human mind. *Evolutionary Psychology*, 3, 59-78.
- Freud, S. (1950). *The interpretation of dreams*. New York, NY: Random House.
- Gardner, W. L., Pickett, C. L. & Brewer, M. B. (2000). Social exclusion and selective memory: How the need to belong influences memory of social events. *Personality and Social Psychology Bulletin*, 26 (4), 486-496. [10.1177/0146167200266007](https://doi.org/10.1177/0146167200266007)
- Germain, A., Shear, K. M., Walsh, C., Buysse, D. J., Monk, T. H., Reynolds, C. F., Frank, E. & Silowash, R. (2013). Dream content in complicated grief: A window into loss-related cognitive schemas. *Death Studies*, 37 (3), 269-284. [10.1080/07481187.2011.641138](https://doi.org/10.1080/07481187.2011.641138)
- Hall, C. S. & Van de Castle, R. L. (1966). *The content analysis of dreams*. New York, NY: Appleton-Century-Crofts.
- Hamilton, W. (1964). The genetical evolution of social behaviour. *Journal of Theoretical Biology*, 7 (1), 1-16. [10.1016/0022-5193\(64\)90038-4](https://doi.org/10.1016/0022-5193(64)90038-4)
- Hartmann, E. (1995). Making connections in a safe place: Is dreaming psychotherapy? *Dreaming*, 5 (4), 213-228. [10.1037/h0094437](https://doi.org/10.1037/h0094437)
- (1996). Outline for a theory on the nature and functions of dreaming. *Dreaming*, 6 (2), 147-170. [10.1037/h0094452](https://doi.org/10.1037/h0094452)
- (1998). *Dreams and nightmares: The new theory on the origin and meaning of dreams*. New York, NY: Plenum Press.
- Henrich, J., Heine, S. J. & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33, 61-83. [10.1017/S0140525X0999152X](https://doi.org/10.1017/S0140525X0999152X)
- Heynick, F. (1993). *Language and its disturbances in dreams*. New York, NY: Wiley.
- Hill, K. R., Walker, R. S., Božičević, M., Eder, J., Headland, T., Hewlett, B., Hurtado, M. & Wood, B. (2011). Co-residence patterns in hunter-gatherer societies show

- unique human social structure. *Science*, 331 (6022), 1286-1289. [10.1126/science.1199071](https://doi.org/10.1126/science.1199071)
- Hobson, J. A. (1988). *The dreaming brain*. New York, NY: Basic Books.
- (1997). Dreaming as delirium: A mental status exam of our nightly madness. *Seminars in Neurology*, 17 (2), 121-128. [10.1055/s-2008-1040921](https://doi.org/10.1055/s-2008-1040921)
- (2001). *The dream drugstore: Chemically altered states of consciousness*. Cambridge, MA: MIT Press.
- (2009). REM sleep and dreaming: Towards a theory of protoconsciousness. *Nature Reviews Neuroscience*, 10 (11), 803-813. [10.1038/nrn2716](https://doi.org/10.1038/nrn2716)
- (2011). *Dream life*. Cambridge, MA: MIT Press.
- Hobson, J. A., Pace-Schott, E. F. & Stickgold, R. (2000). Dream science 2000: A response to commentaries on Dreaming and the brain. *Behavioral and Brain Sciences*, 23 (6), 1019-1035. [10.1017/S0140525X00954025](https://doi.org/10.1017/S0140525X00954025)
- Hobson, J. A. & Friston, K. J. (2012). Waking and dreaming consciousness: Neurobiological and functional considerations. *Progress in Neurobiology*, 98 (1), 82-98. [10.1016/j.pneurobio.2012.05.003](https://doi.org/10.1016/j.pneurobio.2012.05.003)
- Hobson, J. A. & Schredl, M. (2011). The continuity and discontinuity between waking and dreaming: A dialogue between Michael Schredl and Allan Hobson concerning the adequacy and completeness of these notions. *International Journal of Dream Research*, 4 (1), 3-7. [10.11588/ijodr.2011.1.9087](https://doi.org/10.11588/ijodr.2011.1.9087)
- Humphrey, N. (2000). Dreaming as play. *Behavioral and Brain Sciences*, 23 (6), 953-953. [10.1017/S0140525X0054026](https://doi.org/10.1017/S0140525X0054026)
- Kahn, D., Pace-Schott, E. & Hobson, J. A. (2002). Emotion and cognition: Feeling and character identification in dreaming. *Consciousness and Cognition*, 11 (1), 34-50. [10.1006/ccog.2001.0537](https://doi.org/10.1006/ccog.2001.0537)
- Kahn, D. & Hobson, J. A. (2005). Theory of mind in dreaming: Awareness of feelings and thoughts of others in dreams. *Dreaming*, 15 (1), 48-57. [10.1037/1541-1559.15.1.48](https://doi.org/10.1037/1541-1559.15.1.48)
- Lara-Carrasco, J., Simard, V., Saint-Onge, K., Lamoureux-Tremblay, V. & Nielsen, T. A. (2013). Maternal representations in the dreams of pregnant women: a prospective comparative study. *Frontiers in Psychology*, 4, 1-13. [10.3389/fpsyg.2013.00551](https://doi.org/10.3389/fpsyg.2013.00551)
- (2014). Disturbed dreaming during the third trimester of pregnancy. *Sleep Medicine*, 15 (6), 694-700. [10.1016/j.sleep.2014.01.026](https://doi.org/10.1016/j.sleep.2014.01.026)
- Leary, M. R., Tambor, E. S., Terdal, S. K. & Downs, D. L. (1995). Self-esteem as an interpersonal monitor: The sociometer hypothesis. *Journal of Personality and Social Psychology*, 68 (3), 518-530. [10.1037/0022-3514.68.3.518](https://doi.org/10.1037/0022-3514.68.3.518)
- Leary, M. R., Haupt, A. L., Strausser, K. S. & Chokel, J. T. (1998). Calibrating the sociometer: The relationship between interpersonal appraisals and state self-esteem. *Journal of Personality and Social Psychology*, 74 (5), 1290-1299. [10.1037/0022-3514.74.5.1290](https://doi.org/10.1037/0022-3514.74.5.1290)
- Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, 64, 354-361.
- Llinás, R. (2001). *I of the vortex: From neurons to self*. Cambridge, MA: MIT Press.
- Llinás, R. R. & Paré, D. (1991). Of dreaming and wakefulness. *Neuroscience*, 44 (3), 521-535. [10.1016/0306-4522\(91\)90075-Y](https://doi.org/10.1016/0306-4522(91)90075-Y)
- Llinás, R. & Ribary, U. (1994). Perception as an oneiric-like state modulated by the senses. In C. Koch & J. L. Davis (Eds.) *Large-scale neuronal theories of the brain* (pp. 111-124). Cambridge, MA: MIT Press.
- McNamara, P. (1996). REM sleep: A social bonding mechanism. *New Ideas in Psychology*, 14 (1), 35-46. [10.1016/0732-118X\(95\)00023-A](https://doi.org/10.1016/0732-118X(95)00023-A)
- (1996). REM sleep: A social bonding mechanism. *New Ideas in Psychology*, 14 (1), 35-46. [10.1016/0732-118X\(95\)00023-A](https://doi.org/10.1016/0732-118X(95)00023-A)
- McNamara, P., McLaren, D., Smith, D., Brown, A. & Stickgold, R. (2005). A “Jekyll and Hyde” within: aggressive versus friendly interactions in REM and non-REM dreams. *Psychological Science*, 16 (2), 130-136. [10.1111/j.0956-7976.2005.00793.x](https://doi.org/10.1111/j.0956-7976.2005.00793.x)
- McNamara, P., McLaren, D., Kowalczyk, S. & Pace-Schott, E. (2007). “Theory of mind” in REM and NREM dreams. In D. Barrett & P. McNamara (Eds.) *The new science of dreaming* (pp. 201-220). Westport, CT: Praeger.
- Mealey, L., Daood, C. & Krage, M. (1996). Enhanced memory for faces of cheaters. *Ethology and Sociobiology*, 17 (2), 119-128. [10.1016/0162-3095\(95\)00131-X](https://doi.org/10.1016/0162-3095(95)00131-X)
- Merei, F. (1994). Social relationships in manifest dream content. *Journal of Russian and East European Psychology*, 32 (1), 46-88. [10.2753/RPO1061-0405320146](https://doi.org/10.2753/RPO1061-0405320146)
- Metzinger, T. (2003). *Being no one: The self-model theory of subjectivity*. Cambridge, MA: MIT Press.
- Nielsen, T. A. (2000). A review of mentation in REM and NREM sleep: “Covert” REM sleep as a possible reconciliation of two opposing models. *Behavioral and Brain Sciences*, 23 (6), 851-866. [10.1017/S0140525X0000399X](https://doi.org/10.1017/S0140525X0000399X)
- (2010). Dream analysis and classification: The reality simulation perspective. In M. Kryeger, T. Roth & W. C. Dement (Eds.) *Principles and Practice of*

- Sleep Medicine* (pp. 595-603). New York, NY: Elsevier.
- Nielsen, T. A. & Germain, A. (2000). Post-traumatic nightmares as a dysfunctional state. *Behavioral and Brain Sciences*, 23 (6), 978-979. [10.1017/S0140525X0070402X](https://doi.org/10.1017/S0140525X0070402X)
- Nielsen, T. A. & Lara-Carrasco, J. (2007). Nightmares, dreaming, and emotion regulation. In D. Barrett & P. McNamara (Eds.) *The new science of dreaming* (pp. 253-284). Westport, CT: Praeger.
- Nielsen, T. A. & Levin, R. (2007). Nightmares: A new neurocognitive model. *Sleep Medicine Reviews*, 11 (4), 295-310. [10.1016/j.smrv.2007.03.004](https://doi.org/10.1016/j.smrv.2007.03.004)
- Nielsen, T. A. & Powell, R. A. (1989). The “dream-lag” effect: A 6-day temporal delay in dream content incorporation. *Psychiatric Journal of the University of Ottawa*, 14 (4), 561-565.
- Oberman, L. M. & Ramachandran, V. S. (2007). The simulating social mind: The role of the mirror neuron system and simulation in the social and communicative deficits of autism spectrum disorders. *Psychological Bulletin*, 133 (2), 310-327. [10.1037/0033-2909.133.2.310](https://doi.org/10.1037/0033-2909.133.2.310)
- Paul, F. & Schredl, M. (2012). Male-female ratio in waking-life contacts and dream characters. *International Journal of Dream Research*, 5 (2), 119-124. [10.11588/ijodr.2012.2.9406](https://doi.org/10.11588/ijodr.2012.2.9406)
- Pickett, C. L., Gardner, W. L. & Knowles, M. (2004). Getting a cue: The need to belong and enhanced sensitivity to social cues. *Personality and Social Psychology Bulletin*, 30 (9), 1095-1107. [10.1177/0146167203262085](https://doi.org/10.1177/0146167203262085)
- Rechtschaffen, A. & Buchignani, C. (1992). The visual appearance of dreams. In J. S. Antrobus & M. Bertini (Eds.) *The neuropsychology of sleep and dreaming* (pp. 143-155). Hillsdale, NJ: Lawrence Erlbaum.
- Revonsuo, A. (1995). Consciousness, dreams, and virtual realities. *Philosophical Psychology*, 8 (1), 35-58. [10.1080/095115089508573144](https://doi.org/10.1080/095115089508573144)
- (2000). The reinterpretation of dreams: An evolutionary hypothesis of the function of dreaming. *Behavioral and Brain Sciences*, 23 (6), 877-901. [10.1017/S0140525X00004015](https://doi.org/10.1017/S0140525X00004015)
- (2005). The self in dreams. In T. E. Feinberg & J. P. Keenan (Eds.) *The lost self: Pathologies of the brain and mind* (pp. 206-219). New York, NY: Oxford University Press.
- (2006). *Inner presence: Consciousness as a biological phenomenon*. Cambridge, MA: MIT Press.
- Revonsuo, A. & Salmivalli, C. (1995). A content analysis of bizarre elements in dreams. *Dreaming*, 5 (3), 169-187. [10.1037/h0094433](https://doi.org/10.1037/h0094433)
- Revonsuo, A. & Tarkko, K. (2002). Binding in dreams. *Journal of Consciousness Studies*, 9 (7), 3-24.
- Robert, G. & Zadra, A. (2014). Thematic and content analysis of idiopathic nightmares and bad dreams. *Sleep*, 37 (2), 409-417. [10.5665/sleep.3426](https://doi.org/10.5665/sleep.3426)
- Roll, S. & Millen, L. (1979). The friend as represented in the dreams of late adolescents: Friendship without rose-coloured glasses. *Adolescence*, 14 (54), 255-275.
- Schredl, M. (2011). Dreams of a romantic partner in a dream series: Comparing relationship periods with periods of being separated. *International Journal of Dream Research*, 4 (2), 127-131. [10.11588/ijodr.2011.2.9150](https://doi.org/10.11588/ijodr.2011.2.9150)
- (2012). Old school friends: Former social relationship patterns in a long dream series. *International Journal of Dream Research*, 5 (2), 143-147. [10.11588/ijodr.2012.2.9432](https://doi.org/10.11588/ijodr.2012.2.9432)
- (2013). Dreams of core family members in a long dream series. *International Journal of Dream Research*, 6 (2), 114-118. [10.11588/ijodr.2013.2.11055](https://doi.org/10.11588/ijodr.2013.2.11055)
- Schredl, M. & Hofmann, F. (2003). Continuity between waking activities and dream activities. *Consciousness and Cognition*, 12 (2), 298-308. [10.1016/S1053-8100\(02\)00072-7](https://doi.org/10.1016/S1053-8100(02)00072-7)
- Schweickert, R. (2007). Social networks of characters in dreams. In D. Barrett & P. McNamara (Eds.) *The new science of dreaming*. Westport, CT: Praeger.
- Selتمان, D. & Drigotas, S. (2009). Attachment styles and emotional content, stress, and conflict in dreams of romantic partners. *Dreaming*, 19 (3), 135-151. [10.1037/a0017087](https://doi.org/10.1037/a0017087)
- Selتمان, D. F., Apetroaia, A. I., Riela, S. & Aron, A. (2014). Dreaming of you: Behavior and emotion in dreams of significant others predict subsequent relational behaviour. *Social Psychological and Personality Science*, 5 (1), 111-118. [10.1177/1948550613486678](https://doi.org/10.1177/1948550613486678)
- Strauch, I. & Meier, B. (1996). *In search of dreams: Results of experimental dream research*. New York, NY: SUNY Press.
- Stumbrys, T., Erlacher, D. & Schmidt, S. (2011). Lucid dream mathematics: An explorative online study of arithmetic abilities of dream characters. *International Journal of Dream Research*, 4 (1), 35-40. [10.11588/ijodr.2011.1.9079](https://doi.org/10.11588/ijodr.2011.1.9079)
- Sutcliffe, A., Dunbar, R., Binder, J. & Arrow, H. (2012). Relationships and the social brain: Integrating psychological and evolutionary perspectives. *British Journal of Psychology*, 103 (2), 149-168. [10.1111/j.2044-8295.2011.02061.x](https://doi.org/10.1111/j.2044-8295.2011.02061.x)

- Tholey, P. (1989). Consciousness and abilities of dream characters observed during lucid dreaming. *Perceptual and Motor Skills*, 68 (2), 567-578.
[10.2466/pms.1989.68.2.567](https://doi.org/10.2466/pms.1989.68.2.567)
- Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, 46 (1), 35-57.
- Uguccioni, G., Golmard, J. L., de Fontréaux, A. N., Leu-Semenescu, S., Brion, A. & Arnulf, I. (2013). Fight or flight? Dream content during sleepwalking/sleep terrors vs. rapid eye movement sleep behavior disorder. *Sleep Medicine*, 14 (5), 391-398. [10.1016/j.sleep.2013.01.014](https://doi.org/10.1016/j.sleep.2013.01.014)
- Valli, K., Revonsuo, A., Pälkä, O. & Punamäki, R.-L. (2006). The effect of trauma on dream content: A field study of Palestinian children. *Dreaming*, 16 (2), 63-87.
[10.1037/1053-0797.16.2.63](https://doi.org/10.1037/1053-0797.16.2.63)
- Valli, K., Strandholm, T., Sillanmäki, L. & Revonsuo, A. (2008). Dreams are more negative than real life: Implications for the function of dreaming. *Cognition and Emotion*, 22 (5), 833-861. [10.1080/02699930701541591](https://doi.org/10.1080/02699930701541591)
- Valli, K. & Revonsuo, A. (2006). Recurrent dreams: Recurring threat simulations? *Consciousness and Cognition*, 15 (2), 470-474. [10.1016/concog.2005.05.2001](https://doi.org/10.1016/concog.2005.05.2001)
- (2009). The threat simulation theory in light of recent empirical evidence: A review. *American Journal of Psychology*, 122 (1), 17-38.
- Windt, J. M. (2010). The immersive spatiotemporal hallucination model of dreaming. *Phenomenology and Cognitive Science*, 9 (2), 295-316.
[10.1007/s11097-010-9163-1](https://doi.org/10.1007/s11097-010-9163-1)
- Windt, J. M. & Metzinger, T. (2007). The philosophy of dreaming and self-consciousness: What happens to the experiential subject during the dream state? In D. Barrett & P. McNamara (Eds.) *The new science of dreaming* (pp. 193-247). Westport, CT: Praeger.
- Zadra, A., Desjardins, S. & Marcotte, E. (2006). Evolutionary function of dreams: A test of the threat simulation theory in recurrent dreams. *Consciousness and Cognition*, 15, 450-463. [10.1016/j.concog.2005.02.002](https://doi.org/10.1016/j.concog.2005.02.002)

The Multifunctionality of Dreaming and the Oblivious Avatar

A Commentary on Revonsuo & Colleagues

Martin Dresler

Sleep and dreaming do not serve a single biological function, but are multifunctional. Their functions include memory consolidation and integration, emotion regulation, creativity and problem solving, and preparation for waking life. One promising level of description is that of dreaming as a virtual reality: The dreamer interacts with a simulated environment including other simulated avatars. While dreaming can be considered a multifunctional general reality simulator, the threat simulation and social simulation functions of dreaming are unique among other dream functions in their ability to explain a striking feature of dream phenomenology: obliviousness towards the true state of mind.

Keywords

Avatars | Creativity | Dream | Dreaming | Emotion regulation | Function | Lucid dreaming | Memory | Multifunctional general reality simulator | REM sleep | Simulation | Sleep | Social simulation theory | Threat simulation theory | Virtual reality

Commentator

[Martin Dresler](#)

martin.dresler@donders.ru.nl

Radboud Universiteit Medical Center
Nijmegen, Netherlands

Target Authors

[Antti Revonsuo](#)

antti.revonsuo@utu.fi

Högskolan i Skövde, Skövde, Sweden
Turun yliopisto, Turku, Finland

[Jarno Tuominen](#)

jarno.tuominen@utu.fi

Turun yliopisto
Turku, Finland

[Katja Valli](#)

katval@utu.fi

Turun yliopisto, Turku, Finland
Högskolan i Skövde, Skövde, Sweden

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

Sleep is an almost ubiquitous phenomenon within the animal kingdom, existing in all higher and many lower species. The specific function of sleep, however, is still an enigma:

sleep helps an organism to save energy through extended periods of inactivity, yet at the same time leaves it in a potentially dangerous state of non-responsiveness. While several possible functions of sleep have been discussed in recent years (Frank 2006; Vassalli & Dijk 2009), the function of dreaming might be seen as an even bigger mystery: the hyper-realistic imagery experienced during dreaming does not inform the organism about its current environment, and the virtual motor activity processed in interaction with these hallucinations is not executed to affect the external world—or even worse, in pathological conditions like REM sleep behavior disorder it is, thereby threatening the health of the dreamer and his bed partner. After awakening from a dream, the often emotionally-toned preoccupation with the dream narrative can confuse the dreamer and distract him from potentially dangerous conditions in the real world.

An increasingly widespread idea is that the function of dreaming consists in the simulation of waking life. In a variation of their threat simulation theory (TST; Revonsuo 1995, 2000), Revonsuo et al. (this collection) now propose a social simulation theory of dreaming (SST), according to which dream function could best be characterized as simulating social reality. Considering the social nature of most of our dreams, SST is an intuitively plausible approach, and Revonsuo et al. review a number of studies that provide support for SST. Nevertheless, several questions remain to be clarified: is the prime function of dreaming threat simulation or social simulation—or something completely different? What is the relationship between the various proposed functions of sleep and dreaming, including TST and SST? If the TST and SST turn out not to be the sole or even prime functions of dreaming, do they nevertheless provide unique insights into the function of dreaming?

In this commentary, I shall review several widely propagated functions of sleep and dreaming. I shall then compare these functions with the social and threat simulation functions of dreaming, and finally discuss why and in which regard these two functions might be special. I shall argue that the merit of TST and

SST is not the conclusive explanation of the function of dreaming—which I consider a multifunctional state—but that they are the only candidates among the variety of dream functions that are capable of explaining a striking feature of most dreams: obliviousness towards the current state of mind.

2 Sleep physiology and the function of dreaming

When speculating about the function of dreaming, some clarifications about the level of explanation are necessary. By definition (e.g., Windt 2010), dreaming is a phenomenon occurring during sleep. In an account of biological realism (Revonsuo 2006), the function of dreaming cannot be discussed independently from the neurophysiology of sleep. Even if the phenomenology of dreaming serves a function that can be conceptually (and maybe evolutionarily) differentiated from the original function realized by its physiological correlates, this function is not independent from the neurophysiology of sleep and its specific functions: if the neurophysiological functions change their mechanisms, this would also affect the phenomenological aspects of dreaming—philosophically speaking, phenomenal properties of dreaming supervene on neurophysiological properties of sleep. However, neither can the function of dreaming be equated with the function of sleep, since there are functions of sleep for which it is rather unlikely that any phenomenological aspects play a role, e.g., myelin sheath proliferation (Bellesi et al. 2013); synaptic downscaling (Tononi & Cirelli 2006); metabolite clearance (Xie et al. 2013); or general metabolic (Morselli et al. 2012) and immunological functions (Besedovsky et al. 2012). There are also functions of sleep that might be described conceptually without referring to phenomenal aspects, but in fact happen to be biologically associated with dream mentation, e.g., physiological microprocesses underlying memory consolidation (see below). And in these cases, one can differentiate dream phenomenology and sleep physiology on a conceptual, but not biological level—unless one adopts a radically dualistic approach, that is. Hence, speaking of the

function of dreaming—in contrast to the function of sleep more generally—always implies both phenomenological and physiological aspects.

When considering the neurophysiology of dreaming, coarse sleep stages as defined by classical polysomnography have been the prime targets of investigation. Among these, REM sleep harbors the most prototypical dreams, with a story-like dream narrative including interactive visuomotor hallucinations and often intense emotions. In addition, REM sleep dreams can be most elegantly related to their neurophysiological correlates (Hobson & Pace-Schott 2002). Nevertheless, dream-like mentation can be found in all sleep stages (Nielsen 2000), and hence also the neurophysiology of other sleep stages has to be taken into account when investigating the function of dreaming. In conclusion, when speculating about the function of dreaming, all those REM and NREM sleep functions have to be considered that can reasonably be expected to be associated with phenomenal aspects. In the following, I will highlight four clusters of such sleep functions.

3 Dream function 1: Memory consolidation and integration

In recent years, the most widely discussed function of sleep and dreaming concerns the consolidation of declarative memory, including semantic, episodic, and autobiographical information; and procedural memory including perceptual and motor skills (Rasch & Born 2013). In particular the role of REM sleep in memory consolidation has been studied for several decades. While many studies from the 1970s have been criticized for being heavily confounded by too stressful REM sleep deprivation procedures (Horne & McGrath 1984), research in the 1990s raised interest in the role of REM sleep for memory consolidation: Karni (1994) demonstrated that a basic visual discrimination task improved after a normal night's sleep, but not after selective REM sleep deprivation. Following this, a leading research aim in the field has been to identify which memory systems benefit from which sleep stages: it was demonstrated that

early deep sleep benefits declarative memories, while late REM-rich sleep supports procedural skills (Plihal & Born 1997). Further support for the role of REM sleep in procedural memory consolidation came from studies showing that REM sleep intensity (total number of REMs and REM densities) increased following procedural-task acquisition (Smith et al. 2004) and improvements in procedural memory performance after a night of sleep were proportional to time spent in REM sleep (Fischer et al. 2002). Moreover, brain areas activated during a procedural learning task were more active during REM sleep in subjects who were trained at the task (Maquet et al. 2000; Peigneux et al. 2003).

More recent studies, however, speak against a prominent role of REM sleep in the consolidation of procedural motor skills or other forms of non-emotional memories, and instead emphasize non-REM sleep processes (Genzel et al. 2014). On the neurophysiological level, it has been suggested that dreaming represents the phenomenological reflection of a neural replay of activation patterns associated with recent learning experiences (Wilson & McNaughton 1994; Wamsley & Stickgold 2011; Wamsley 2014). Although memory reactivations have been observed in REM sleep as well (Louie & Wilson 2001), the most advanced models of sleep-related memory consolidation propose that neural replay is orchestrated by an interaction of non-REM sleep microprocesses, including slow oscillations and sleep spindles (Genzel et al. 2014).

Events and episodes from waking life are sometimes incorporated into dreams, either as classical day-residues the following night or after a “dream lag” of about 5–7 days (Nielsen & Powell 1989; Nielsen et al. 2004). Supporting the idea that such dream incorporations reflect processes of memory consolidation, items that were incorporated into dreams have been observed to lead to better memory retention (de Koninck et al. 1990; Cipolli et al. 2004). While an actual episodic replay of waking events was found in no more than 1–2% of the dream reports (Fosse et al. 2003), with NREM-sleep dreams appearing to include more identifiable episodic memory sources than REM-sleep

dreams (Baylor & Cavallero 2001), it has been suggested that particularly engaging learning experiences have a more robust influence on dream content relative to more passive experiences (Wamsley 2014).

In contrast to recent episodes, incorporations of autobiographical memory features could be identified in the majority of dreams (Malinowski & Horton 2014). This suggests that dreaming might serve to assimilate recent memory fragments into autobiographical memory schemas and thus supports autobiographical self-model maintenance (Metzinger 2013). For semantic memories, evidence of a relationship between dreaming and neural memory reactivations stems from studies of declarative memory that present memory cues during sleep: these cues, when associated with the pre-sleep learning session, induce associated dream imagery (Schredl et al. 2014) and enhance post-sleep memory retrieval (Rasch et al. 2007). For procedural memories, learning of an engaging visuomotor task led to integration of task-related imagery into dream-like activity during non-REM sleep (Wamsley et al. 2010a), and such dream-incorporations of recent learning experiences were associated with later memory performance (Wamsley et al. 2010b). This memory-enhancing re-experience reminds us of motor imagery training during wakefulness, which has been repeatedly demonstrated to improve motor skills (Driskell et al. 1994; Schuster et al. 2011).

Recently it has been suggested that instead of consolidating memories, REM sleep serves as a state of elaborative (re-)encoding, during which the hippocampus integrates recent episodic memory fragments into remote episodic memories (Llewellyn 2013). It has been proposed that this process relies upon principles that also underlie the mnemonic encoding strategies of ancient orators, such as vivid, complex and often bizarre associative imagery, narratives with embodiment of oneself, and associations with known locations, later serving as retrieval cues. Subjectively, this process would be experienced as the typical dream mentation with its hyper-associative and bizarre imagery. However, despite being intuitively appealing,

several theoretical considerations and empirical findings are inconsistent with the idea of mnemonic encoding strategies acting during dreaming (Dresler & Konrad 2013).

To sum up, a first important function of sleep and dreaming is memory consolidation and integration, including the rehearsal of procedural motor skills, replay of episodic and semantic memories, and integration of memory episodes into autobiographical memory schemas.

4 Dream function 2: Emotion regulation

Converging evidence suggests that the regulation of emotional processes is an important function of sleep and dreaming. Early content analyses of REM sleep dreams showed that many dreams are highly emotional, with unpleasant emotions prevailing (Hall & Van de Castle 1966; Snyder 1970). This is in line with neuroimaging studies of REM sleep, demonstrating that neural areas involved in emotion regulation like the amygdala, medial prefrontal cortex, and anterior cingulate cortex are highly activated during REM sleep (Nir & Tononi 2010). Several REM-sleep characteristics differ between healthy subjects scoring low in depression scales and those with higher but still sub-clinical depression scores (Cartwright et al. 1998). After highly emotional life events, REM sleep changes can be observed in those subjects that react with symptoms of depression (Cartwright 1983), and dreams of depressed subjects differ from patients in remission (Cartwright et al. 2006). Likewise, in depressed patients the distribution of rapid eye movements in REM sleep differs in nights after which mood is estimated better than in the preceding evening compared to nights after which mood is unchanged (Indursky & Rotenberg 1998). It was therefore proposed that REM sleep dreaming serves as a mood regulation system and that a disturbance of this process might play a role in the development of affective disorders (Cartwright 2011). Changes in REM sleep are symptomatic of affective disorders and the sleep-memory relationship is altered in these diseases (Dresler et al. 2014). In healthy subjects, the consolidation of emotional texts

(Wagner et al. 2001) or pictures (Hu et al. 2006; Nishida et al. 2009) is enhanced through REM sleep, an effect that has been shown to last for several years (Wagner et al. 2006).

While at first sight it might look as if REM sleep unequivocally strengthens emotional memory processes, some studies suggest a more complex picture: referring to the fact that emotional experiences are remembered better than neutral ones, however their emotional tone during retrieval decreases with time, it was proposed that REM sleep serves an emotional decoupling function: we sleep to remember emotionally-tagged information yet at the same time to forget the associated emotional tone (Walker & van der Helm 2009). While some studies support this model (Hu et al. 2006; Nishida et al. 2009), others suggest that the affective tone of emotional memories is preserved rather than reduced during REM sleep (Groch et al. 2013).

Besides negative emotions, sleep and dreaming have also been associated with positive affects. Recent dream report analyses suggest that positive emotions in dreams have been underestimated in previous studies and might be even more common than negative emotions (Malcolm-Smith et al. 2012; Sikka et al. 2014). In addition, the processing of reward has been associated with REM sleep and dreaming. For example, the expectancy of a reward enhances memory consolidation processes during sleep (Fischer & Born 2009), and reactivations of neural activity related to a reward-searching task have been observed in reward-related brain regions such as the ventral striatum during sleep (Pennartz et al. 2004). Instead of a simulation of purely aversive content such as threats, according to this account sleep favors the activation of representations of high emotional and motivational relevance in general (Perogamvros & Schwartz 2012, 2014).

In summary, a second important function of sleep and dreaming is the regulation of emotions, including both an enhancement of emotionally-tagged information and a decoupling of this information from its associated emotional tone.

5 Dream function 3: Creativity and problem solving

Anecdotal reports on scientific discovery, inventive originality, and artistic productivity suggest that creativity can be triggered or enhanced by sleeping and dreaming. Several studies confirm these anecdotes, showing that sleep promotes creative problem-solving compared to wakefulness. For example, when subjects performed a cognitive task that could be solved much faster through applying a hidden rule, after a night of sleep more than twice as many subjects gained insight into the hidden rule as in a control group staying awake (Wagner et al. 2004). Similarly, subjects benefited in a creativity task from an afternoon nap but not from staying awake (Cai et al. 2009; Bejjamini et al. 2014), and the likelihood of solving a problem encountered before sleep can be increased by cued reactivations during sleep (Ritter et al. 2012).

According to the classical stage model of creativity, creative insights may be described by a process consisting of several stages, of which the incubation phase appears to be most intimately associated with sleep and dreaming (Dresler 2011, 2012; Ritter & Dijksterhuis 2014). The most common psychological approaches support this view: psychoanalytical models of creativity emphasize the primary process concept, which denotes free-associative and dream-like thinking, compared to the more rational and analytical secondary-process thinking (Kris 1952). Cognitive models propose that a state of defocused attention facilitates creativity (Mendelsohn 1976)—creative individuals seem to have less narrowly-focused attention than uncreative ones, which leads to unorthodox connections of remote ideas that might eventually lead to creative cognitions. In a similar vein, creative individuals are thought to have relatively flat association hierarchies (i.e., more, yet weaker associations between cognitive elements), which accounts for the ability to make remote associations; whereas uncreative individuals are thought to have relatively steep association hierarchies (Mednick 1962). Physiological models emphasize the level of cortical arousal as an important variable influencing cre-

ativity: both a lower level of cortical arousal—particularly in the prefrontal cortex—and a higher variability in cortical arousal levels are expected in creative compared to uncreative individuals, depending on specific phases of the creative process (Martindale 1999). In addition, low levels of norepinephrine are thought to facilitate creativity, shifting the brain toward intrinsic neuronal activation with an increase in the size of distributed concept representations and co-activation across modular networks (Heilman et al. 2003). The prefrontal cortex seems to be of particular importance for creative processes; however there is evidence that both prefrontal activation and prefrontal deactivation facilitate creativity—maybe depending on the specific phase of the creative process. Brain areas showing selective activation for insight events are—besides the prefrontal cortex—the visual cortices, the hippocampus, and in particular the anterior cingulate cortex, which is thought to be involved in breaking the impasse that marks the critical step of insight into a problem (Dietrich & Kanso 2010).

Both theoretical models and empirical neuroscience of creativity suggest that sleep and dreaming provide an ideal environment for creative incubation: primary-process thinking is explicitly conceptualized as dream-like, and the hyper-associative nature of dreams can be considered a prime example of a flat associative hierarchy. Defocused attention is a phenomenal feature of most dreams, physiologically probably caused by prefrontal cortex deactivation. And daydreaming has the potential to increase creativity (Lewin 1989), while the level of engagement in such mind-wandering in contrast to explicitly directed thoughts is associated with creative performance (Baird et al. 2012). The sleep cycle provides the brain with highly alternating arousal levels, and the chaotic activation of the cortex in REM sleep through brain stem regions in absence of external sense data leads to a much more radical renunciation of unsuccessful problem solving attempts, leading to co-activations of cognitive data that are highly remote in waking life (Kahn et al. 2002a). These co-activations, woven into a dream narrative in a self-organizing manner, repeatedly receive further

innervations by the brainstem, leading to bizarre sequences of loosely associated dream topics that might eventually activate particular problem-relevant cognitions or creative cognitions in general (Hobson & Wohl 2005). In addition, in REM sleep, which is characterized by low levels of norepinephrine, visual cortices, the hippocampus, and the anterior cingulate cortex have all been shown to be strongly activated, potentially facilitating insight events. In conclusion, the phenomenological and neural correlates of sleeping and dreaming provide ideal conditions for the genesis of creative ideas and insights.

In summary, a third important function of sleep and dreaming is the association of remote cognitive elements in order to facilitate creativity and problem solving.

6 Dream function 4: Preparation and simulation of waking life

Consolidation, integration, regulation, and re-evaluation of acquired information during sleep prepare the organism for its waking life. However, such processes do not necessarily need to be purely reactive, depending solely on the experiences of the preceding day: several authors propose that a major function of sleep and dreaming might include primarily preparational mechanisms. Since REM sleep dominates sleep more during early developmental periods in comparison to later in life, some researchers have argued that REM sleep plays a role in early brain maturation (Roffwarg et al. 1966; Marks et al. 1995; Mirmiran 1995); however, also a life-long preparational function of REM sleep has been proposed. One of the first approaches in this direction was offered by Jouvett (1979), who combined the brain maturation hypothesis with a metaphor offered by Dewan (1970), in which he claims that the brain is a computer that is programmed during REM sleep—suggesting that innate behaviors are rehearsed during REM-sleep dreaming in order to prepare the organism for their application in waking life. Jouvett later revised his approach, assuming that REM sleep constitutes an iterative genetic programming that helps to maintain

the process of psychological individuation (Jouvet 1998). In a similar vein, Hobson (2009) proposed that REM sleep may constitute a “protoconscious” state, preparing the organism for waking conscious experiences. The development of consciousness during ontogenetic development in this view is a gradual and lifelong process, building on the more primitive innate virtual reality generator, which is phenomenally experienced as dreaming. With the recent integration of Friston’s (2010) predictive coding approach into this theory, the brain is thought to run a virtual world model (see also Revonsuo 1995, 2006; Metzinger 2003) that is continuously updated by processing prediction errors during wakefulness. Freed from external sensory constraints, processing of prediction errors in the dreaming brain actively refines intermediate hierarchy levels of the virtual world model. Dreaming thereby minimizes internal model complexity in order to generate more efficient predictions during subsequent wakefulness (Hobson & Friston 2012; Hobson et al. 2014).

One of the first and today the most widely discussed preparational approach is based on the observation that during dreaming particularly threatening experiences are overrepresented: the Threat Simulation Theory (TST) proposes that one function of sleep is to simulate threatening events, and to rehearse threat perception and threat avoidance (Revonsuo 1995, 2000). Such a mechanism of simulating the threats of waking life over and over again in various combinations would be valuable for the development and maintenance of threat-avoidance skills. Several empirical studies support TST (Revonsuo 2006; Valli & Revonsuo 2009), however some inconstant findings have been reported (Zadra et al. 2006; Malcolm-Smith et al. 2008, 2012). In a variation of TST, Revonsuo et al. (this collection) propose the Social Simulation Theory (SST), according to which the function of dreaming consists in the simulating of “the social skills, bonds, interactions and networks that we engage in during our waking lives”. The SST aims to predict and explain the simulations of social interaction of dream avatars that happen outside threatening events in dreams. Like the TST, predictions of the SST

are supported by a number of studies, but face inconsistent data (Revonsuo et al. this collection).

On a neurobiological level, empirical support for simulation theories of dreaming comes from a recent study demonstrating that the ventromedial prefrontal cortex subserves the simulation and evaluation of possible future experiences, integrating arbitrary combinations of knowledge structures to simulate the emergent affective quality that a possible future episode may hold (Benoit et al. 2014). As the ventromedial prefrontal cortex is known to be activated in REM sleep (Nir & Tononi 2010), this mechanism might also underlie episodes of reality simulation during dreaming. Further neurobiological support for the preparational role of sleep comes from recent research demonstrating a neural “preplay” of future learning-related place-cell sequences in the hippocampus (Dragoi & Tonegawa 2011, 2013). In contrast to the intuitive view that such activation patterns are established for the first time during a novel experience, according to these findings the specific temporal firing sequence during learning seems rather to be selected from a larger repertoire of preexisting activation patterns, thus suggesting that sleep plays a role not only in the subsequent consolidation, but also in the preceding preparation for new experiences. It has been demonstrated that sleep preceding the learning experience indeed influences memory acquisition during the following day (van der Werf et al. 2009). Interestingly, support for the hypothesis that sleep mentation constitutes a virtual reality model preparing for waking life comes also from research outside of sleep neuroscience: approaches probing artificial intelligence demonstrate that robots perform better in navigational tasks if they create and update models of their own structure and actions during a state of motoric inactivity (Bongard et al. 2006). Not surprisingly, this process of evaluation and simulation of prior and future actions was interpreted as dream-like (Adami 2006).

In summary, a fourth important function of sleep and dreaming is preparation for waking life. This includes proposals of REM sleep as an iterative genetic programming system, dreaming

as a state of protoconsciousness and virtual world model optimization, and dreaming as a simulation of threats (TST) and social interactions (SST).

7 The multifunctionality of dreaming

Numerous suggestions for solving the mystery of sleep and dream function can be found in the literature. In the previous sections I have reviewed four clusters of proposed functions of sleep and dreaming: 1) consolidation of recently acquired memories, including procedural motor skill rehearsal, replay of recently acquired memories, and integration of memory episodes into autobiographical memory schemas; 2) emotion regulation, including both an enhancement of emotionally-tagged information and a decoupling of this information from its associated emotional tone; 3) creativity and problem solving; and (4) preparation and simulation of waking life, including iterative genetic programming, virtual world model optimization, the simulation of threats (TST), and the simulation of social interactions (SST). The question thus remains what the real or primary function of sleep and dreaming is—and what the relationship between the different candidates might be. SST aims to independently cover the social simulations that fall outside the scope of TST, thereby describing an “original evolutionary function of dreams alongside with the threat simulation function of dreaming” (Revonsuo et al. [this collection](#)).

The concept of evolutionary function has been one of the main topics in the philosophy of biology (Mahner & Bunge 2000) and philosophy of mind (Millikan 1984; Neander 1991). Several notions of biological functions exist (Wouters 2003); however a general idea is that the biological function of a trait is determined by its contribution to evolutionary fitness (Walsh & Ariew 1996). Darwin (1871) differentiated between selection occurring as a consequence of ecological factors that directly threaten the organism’s survival, such as predators or other potentially life-threatening dangers of nature, and interactions with members of the same species in order to compete for mating partners.

Both principles, dubbed natural and sexual selection respectively, eventually determine reproductive success as the ultimate decision points for selection. In contemporary accounts, sexual selection was generalized to the concept of social selection, of which the former is considered a subtype (Lyon & Montgomerie 2012; West-Eberhard 2014). The concept of runaway selection, famously illustrated by the evolution of the peacock’s tail, was thought to also be applicable to the evolution of social skills in higher animals, eventually leading to the development of theory of mind, language, dance, or artistic creativity in humans (Flinn & Alexander 2007). This process of an arms race of social skills would require increasing cognitive capacity—and in fact, at least in primates, relative brain size has been related to social group size (Dunbar 1992; Dunbar & Shultz 2007).

It is tempting to associate natural and social selection as the main principles of evolution with TST and SST, respectively. This interpretation would strongly support TST and SST, as it would equate the function of dreaming with two main principles of evolution in general. In this broad sense, however, certain attributes like learning capacity or motor skills increase fitness in terms of natural selection, but do not necessarily serve to help us avoid direct threats. Likewise, certain attributes such as emotion regulation or artistic creativity increase fitness in terms of social selection, but are not necessarily themselves social in a strict sense. Ultimately, of course, all these functions serve reproductive success—however, if any skill ultimately helping us to acquire sexual partners is interpreted as social and any possible obstacle to reproduction is interpreted as a threat, then TST and SST would be trivial, as a biological function is by definition one that supports reproductive success. In contrast, if TST and SST are interpreted in a more narrow, non-trivial way, there is ample space in dreams for further functions: consolidation of navigational information acquired during exploration; rehearsal of a recently learned motor sequence; facilitation of a behavior recently rewarded with food; incubated creative insight into the solution of a recent unsuccessful attempt to build a helpful

tool; refinement of the discriminative skills regarding recently perceived pattern, etc.—all these potential benefits of sleep and dreaming increase inclusive fitness of the individual, but do not directly refer to the simulation of threats or social interactions.

This problem can further be illustrated by Revonsuo's (1995, 2006) approach, where he considers any phenomenal experience as a virtual world model: what is the function of waking consciousness, threat avoidance, or social interaction? Both threat avoidance and social interaction, of course—and many others. That this rather uninformative answer can also be transferred back to the function of dreaming might be illustrated with another ubiquitous example of simulation: in child's play, simulation of real life and the practice of skills needed therein is considered one of the main functions—play allows children to simulate coping with threats in a safe environment, and to develop the social skills needed later in life (Mellou 1994; Pellegrini & Bjorklund 2004). However, these aspects, while important, are not the only functions of play—it also offers the rehearsal of motor and sensory skills, training in predatory behavior, and general intellectual development. Hence, child's play can be considered multifunctional, as can waking or dreaming consciousness.

Segmentation of reality (including dream reality) is possible along numerous lines. In a sense, TST and SST could be interpreted as expressing two orthogonal dimensions of dream space: a security dimension with the directions threat vs. safety, and a sociality dimension with the directions social vs. individual. Dreamed accidents or natural disasters would be characterized by low security and sociality, dreamed experience of bullying by high sociality and low security, and dreamed bonding by high sociality and security, etc. Threat and social interactions in a narrow sense are important aspects both of waking and dreaming life, however they are not the only aspects. Other segmentations are also possible, e.g., by a dimension of motor activity vs. inactivity, or emotional vs. neutral dream content, or a novelty dimension. In the broad sense of natural and social selection, threat and

social interaction would be the two main drivers of evolution, however to the cost that the answer to the question of the function of dreaming becomes a trivial “to support reproductive success”. Of note is that also the other discussed functions might be interpreted within a simulation framework: e.g., simulation visuomotor activity after learning a respective task in the memory function, simulating affective experiences in the emotion regulation function, and simulating problem solving attempts in the creativity function. These different functions are neither mutually exclusive nor strictly independent from each other. In particular the emotion-processing function largely overlaps with both TST and SST—all threats and at least the most important social interactions induce strong emotions, and successful coping with these emotions would be of considerable help when facing threats or social situations. Also other functions of dreaming overlap with TST and SST: consolidation of threat-related information or social gossip improves threat avoidance or social skills, as does creative incubation on threat-related or social problems. On a more abstract level, all these simulations serve the integration of recently experienced information into the behavioral repertoire in order to adapt it to the current waking environment (Hobson et al. 2014).

Identifying the original function of a given trait has proven to be a notoriously difficult issue in the philosophy of biology (Wouters 2013). Dreaming might have originally developed as an epiphenomenon of rather basal neurophysiological sleep functions, and this phenomenological level might eventually have acquired additional functions. Such exaptations (Gould & Vrba 1982) might have been further adapted and in turn developed further neurophysiological exaptations without phenomenological correlates, etc. The original function of dreaming might be unimportant today compared to subsequently evolved functions. Instead of singling out one or two functions of dreaming as original, dreaming might be best seen as a multifunctional general reality simulator, including the simulation of motor skills, emotional processing, problem solving attempts, threats, and social interactions. To follow specific research questions, of

course certain functions still could be highlighted and followed as research heuristics with a given purpose. All functions of sleep and dreaming serve reproductive success ultimately, even though some might be more important than others from a selection point of view. For all dream functions discussed in this chapter, there are convincing supporting but also inconsistent data. The fact that dreaming is not an unselective simulation of the waking world as, e.g., the continuity hypothesis suggests (Schredl & Hofmann 2003), is a sign that some simulation functions might be more important than others. We should note, however, that quantitative overrepresentation of a specific function does not necessarily prove the primacy of this function: different functions might rely on different processes with different timescales, with a highly important function potentially requiring only seconds to be processed, while an unimportant function might take hours. In times of sufficient sleep, dream content related to the relatively unimportant function might thus be overrepresented. The relative importance of one function over another might be tested in cases of scarcity of sleep, e.g., under sleep deprivation, when different functions would have to compete for restricted simulation time. Also of interest in this regard is a comparative approach: it has been demonstrated that sleep propensity, and particularly REM sleep, negatively correlates with predatory risk across species (Lima et al. 2005), which would rather speak against TST. Concerning SST, the tendency to sleep in groups has been reported to negatively correlate with sleep time, which, however, has been interpreted either in terms of social sleep being more efficient due to reduced predatory risk, or as more social species sacrificing sleep to service social relationships during wakefulness (Capellini et al. 2008). Against this background, sleep and dreaming pose an optimization problem: how much time is best spent asleep, spent in specific sleep stages, and spent engaging in specific dream mentation in order to optimize the interplay between the different functions of sleep and dreaming? Dreaming as a general reality simulator might dynamically change its functional priorities, favoring one

over the other of its several functions, depending on the current requirements and constraints of the environment.

8 The oblivious avatar

Even though it is likely that no ‘original’ function of dreaming can be acknowledged, but rather a multiplicity of functions depending on specific research questions and segmentations of the dream space, one aspect of dreaming might distinguish TST and SST from other functions of sleep and dreaming, including other simulation functions: obliviousness of the avatar about being in a dream. Impaired insight into the own state of mind is a hallmark of normal dreaming, (Dresler et al. 2015a). The well-known exception of this symptom of most dreams is the case of lucid dreaming (Dresler et al. 2015), which in turn can be used to test whether state obliviousness is indeed a characterizing feature of TST and SST when compared with other dream functions.

There is no obvious reason why obliviousness about the dream state would be necessary for the memory function of sleep and dreaming. For procedural memory consolidation, lucid dreaming has even been suggested as a state that allows for a hyper-realistic mental training of recently learned motor skills (Erlacher & Chapin 2010). Several studies support this idea: lucidly dreamed training of coin tossing (Erlacher & Schredl 2010) or a finger tapping task (Stumbrys et al. 2015) has been demonstrated to be effective, and a considerable number of professional athletes use lucid dreams to practice sports skills, with most of them having the impression that their performance is thereby improved (Erlacher et al. 2011). For the creativity and insight function of sleep and dreaming, obliviousness regarding the current state of mind is no prerequisite, and lucid dreaming has explicitly been suggested and shown to be used as a tool to increase creative processes (Stumbrys & Daniels 2010; Schädlich & Erlacher 2012; Stumbrys et al. 2014). As with non-lucid dreaming, lucid dreaming is associated with defocused attention and flat association hierarchies—lucid dreams have been reported to include

even more uncommon and bizarre elements than non-lucid dreams (McCarley & Hoffman 1981). At the same time, regained reflective capabilities enable the creative dreamer to evaluate new associations and ideas, a step in the phase model of creativity that for non-lucid dreams is reserved for subsequent wakefulness. This mechanism is illustrated by two interesting case studies: Barrett (2001) describes the case of a painter who in his lucid dreams visited galleries, and then searching for interesting motifs to be painted soon after awakening from the lucid dream. A comparable strategy was used by one of our own study participants (Dresler et al. 2011, 2012), a music composer: when he aimed to compose a new piece of music, he turned on a radio in his lucid dreams and changed radio stations until he heard a composition that he considered interesting. He then woke himself up and wrote the new composition down. In line with these data, questionnaire studies reported that frequent lucid dreamers might be more creative than less-frequent lucid dreamers (Blagrove & Hartnell 2000).

For the emotion regulation function of sleep and dreaming the situation is less clear, however here there is also some evidence indicating that obliviousness is not generally necessary: for the case of positive affects, subjects often report that lucid dreams are associated with particularly positive emotions. And for negative affects, the successful use of lucid dreaming as a therapeutic tool in affective disorders indicates that dream lucidity does not interfere with the emotion regulation function of dreaming (Holzinger 2014).

In contrast, for those cases where a general emotion regulation function of dreaming overlaps with the TST, the necessity of staying ignorant about the true state of consciousness becomes obvious: to successfully serve as an authentic simulation of a threat, the dreamer has to take the threat as real and thus be oblivious towards his true state of mind. The cognitive insight that everything encountered consists only of hallucinated dream imagery and thus cannot harm the dreamer in reality immediately takes the sting out of the threatening experience. This mechanism has been successfully utilized

for recurrent nightmares, where lucid dreaming has been demonstrated to be of therapeutic value (Spoormaker et al. 2003, 2006; Dresler et al. 2015; Rak et al. in press). Thus, for the threat simulation function of dreaming, obliviousness regarding the current state of mind is essential.

For SST, several lines of evidence indicate that obliviousness regarding the current state of mind is a prerequisite for social simulation to be effective. During normal dreams, non-self dream characters are attributed with feelings and thoughts just like in waking life (Kahn & Hobson 2005). Being oblivious about the true nature of these dream characters might ensure that non-perfect social simulations are also taken as autonomous agents instead of mere puppets controlled by the dreamer: dream characters are often implausible compared to their real-life waking counterparts (Kahn & Hobson 2003), however, are nevertheless recognized and identified without major puzzlement (Kahn et al. 2000, 2002b). During a lucid dream, implausible dream characters might be treated less seriously by the dreamer, rendering the social simulation much less effective. This is illustrated by a recent study demonstrating that being tickled by an intentionally-controlled non-self dream character during a lucid dream was comparably ineffective as self-tickling during wakefulness, whereas being unexpectedly tickled by another dream character felt more ticklish (Windt et al. 2014). Non-self dream characters lead to different predictions depending on their perceived autonomy, and their respective simulation thus serves different functions. Lucid dreaming frequency correlates with the amount of control over the dream (Wolpin et al. 1992; Stumbrys et al. 2014), implying that frequent lucid dreamers would conceive dream characters as less autonomous than less frequent lucid dreamers. Thus, although non-self dream characters appear to have quasi-independent mental lives during lucid dreams (Tholey 1989), convincing training of social skills would require the dreamer to be oblivious to the fact that dream characters are not real, but hallucinated.

In summary, in contrast to other functions of sleep and dreaming, TST and SST essentially

depend on state obliviousness of the dreamer. State obliviousness in dreaming might therefore be seen as a prime example of an epistemic constraint of phenomenal experience that leads to new and beneficial functional properties (Metzinger 2003). While both TST and SST (and other functions of sleep) might be applicable to humans and other social animals alike, state obliviousness might be a function that specifically developed in humans: it is unlikely that animals without sophisticated language skills possess the ability to reflect on their current state of mind and compare it to alternative mind-states. In turn, such animals do not need a differential mechanism switching state reflectiveness on and off depending on the current vigilance state. Of note, neural correlates of state reflectiveness, i.e. lucid dreaming, strikingly mirror brain differences seen in humans vs. non-human primates (Dresler et al. 2013).

9 Conclusion

Sleep and dreaming do not serve a single biological function, but are multifunctional states. Their functions include memory consolidation and integration, emotion regulation, creativity and problem solving, and preparation for waking life. One promising description level is that of dreaming as a general reality simulator. TST and SST describe two important purposes of simulation, namely successful coping with threats and social interactions. The merit of TST and SST is not so much that they conclusively explain the function of dreaming—although they represent the two classical principles of evolution, natural and social selection, there are also several other sleep and dream functions. TST and SST might be the only candidates among the multiple functions of sleep and dreaming that explain a particularly striking feature of dream phenomenology: dreaming is a remarkably realistic simulation of waking life, with the exception of a complete failure to successfully reflect on the current state of consciousness. Veridical insight into the dream state is biologically possible, as the phenomenon of lucid dreaming demonstrates. The fact that state reflectiveness is nevertheless generally ab-

sent in dreaming—dream lucidity is a rare phenomenon (Schredl & Erlacher 2011), and even during lucid dreams, lucidity lapses are common (Barrett 1992)—, suggests that state obliviousness during dreaming has an important function. As demonstrated here, among the different candidates for explaining the function of dreaming, TST and SST are the only ones that are capable of elucidating this specific function: state obliviousness is necessary for the effective simulation of threats and social interactions.

Even though recent neurobiological research has begun to reveal the neural correlates of state reflectiveness and, by contrast, of state obliviousness (Voss et al. 2009, 2014; Dresler et al. 2012), the specific neural mechanisms preventing the dreaming brain from realizing its full repertoire of cognitive capabilities are still largely unclear. Further research into these mechanisms might enable exciting opportunities for sleep and dream research by revealing simple methods of dream-lucidity induction. However, if such ways to induce a simulated reality under full control of its user become available too easily and broadly, this might also lead to unforeseen problems, as at least two important functions of dreaming—simulation of threats and social interactions—probably cannot be processed without state obliviousness. This proposed necessity generates a testable hypothesis: individuals with very frequent lucid dreams can be expected to differ from the majority of infrequent lucid dreamers in their threat-avoidance and social skills.

References

- Adami, C. (2006). What do robots dream of? *Science*, 314 (5802), 1093-1094. [10.1126/science.1135929](https://doi.org/10.1126/science.1135929)
- Baird, B., Smallwood, J., Mrazek, M. D., Kam, J. W., Franklin, M. S. & Schooler, J. W. (2012). Inspired by distraction: Mind wandering facilitates creative incubation. *Psychological Science*, 23 (10), 1117-1122. [10.1177/0956797612446024](https://doi.org/10.1177/0956797612446024)
- Barrett, D. (1992). Just how lucid are lucid dreams? *Dreaming*, 2 (4), 221-228.
- (2001). *The committee of sleep*. Norwalk, CT: Crown House Publishing.
- Baylor, G. W. & Cavallero, C. (2001). Memory sources associated with REM and NREM dream reports throughout the night: A new look at the data. *Sleep*, 24 (2), 165-170.
- Beijamini, F., Pereira, S. I., Cini, F. A. & Louzada, F. M. (2014). After being challenged by a video game problem, sleep increases the chance to solve it. *PLoS One*, 9 (1), e84342-e84342. [10.1371/journal.pone.0084342](https://doi.org/10.1371/journal.pone.0084342)
- Bellesi, M., Pfister-Genskow, M., Maret, S., Keles, S., Tononi, G. & Cirelli, C. (2013). Effects of sleep and wake on oligodendrocytes and their precursors. *The Journal of Neuroscience*, 33 (36), 14288-300. [10.1523/JNEUROSCI.5102-12.2013](https://doi.org/10.1523/JNEUROSCI.5102-12.2013)
- Benoit, R. G., Szpunar, K. K. & Schacter, D. L. (2014). Ventromedial prefrontal cortex supports affective future simulation by integrating distributed knowledge. *Proceedings of the National Academy of Sciences of the U.S.A.*, 111 (46), 16550-16555. [10.1073/pnas.1419274111](https://doi.org/10.1073/pnas.1419274111)
- Besedovsky, L., Lange, T. & Born, J. (2012). Sleep and immune function. *Pflügers Archiv European Journal of Physiology*, 463 (1), 121-137. [10.1007/s00424-011-1044-0](https://doi.org/10.1007/s00424-011-1044-0)
- Blagrove, M. & Hartnell, S. J. (2000). Lucid dreaming: Associations with internal locus of control, need for cognition and creativity. *Personality and Individual Differences*, 28 (1), 41-47. [10.1016/S0191-8869\(99\)00078-1](https://doi.org/10.1016/S0191-8869(99)00078-1)
- Bongard, J., Zykov, V. & Lipson, H. (2006). Resilient machines through continuous self-modeling. *Science*, 314 (5802), 1118-1121. [10.1126/science.1133687](https://doi.org/10.1126/science.1133687)
- Cai, D. J., Mednick, S. A., Harrison, E. M., Kanady, J. C. & Mednick, S. C. (2009). REM, not incubation, improves creativity by priming associative networks. *Proceedings of the National Academy of Sciences of the U.S.A.*, 106 (25), 10130-10134. [10.1073/pnas.0900271106](https://doi.org/10.1073/pnas.0900271106)
- Capellini, I., Barton, R. A., McNamara, P., Preston, B. T. & Nunn, C. L. (2008). Phylogenetic analysis of the ecology and evolution of mammalian sleep. *Evolution*, 62 (7), 1764-1776. [10.1111/j.1558-5646.2008.00392.x](https://doi.org/10.1111/j.1558-5646.2008.00392.x)
- Cartwright, R. D. (1983). Rapid eye movement sleep characteristics during and after mood-disturbing events. *Archives of General Psychiatry*, 40 (2), 197-201. [10.1001/archpsyc.1983.01790020095009](https://doi.org/10.1001/archpsyc.1983.01790020095009)
- (2011). Dreaming as a mood regulation system. In M. H. Kryger, T. Roth & W. C. Dement (Eds.) *Principles and Practice of Sleep Medicine* (pp. 620-627). St. Louis, MO: Saunders.
- Cartwright, R., Luten, A., Young, M., Mercer, P. & Bears, M. (1998). Role of REM sleep and dream affect in overnight mood regulation: A study of normal volunteers. *Psychiatry Research*, 81 (1), 1-8. [10.1016/S0165-1781\(98\)00089-4](https://doi.org/10.1016/S0165-1781(98)00089-4)
- Cartwright, R., Agargun, M. Y., Kirkby, J. & Friedman, J. K. (2006). Relation of dreams to waking concerns. *Psychiatry Research*, 141 (3), 261-270. [10.1016/j.psychres.2005.05.013](https://doi.org/10.1016/j.psychres.2005.05.013)
- Cipolli, C., Fagioli, I., Mazzetti, M. & Tuoizzi, G. (2004). Incorporation of presleep stimuli into dream contents: Evidence for a consolidation effect on declarative knowledge during REM sleep? *Journal of Sleep Research*, 13 (4), 317-326.
- Darwin, C. (1871). *Sexual selection and the Descent of Man*. London, UK: Murray.
- de Koninck, J., Christ, G., Hébert, G. & Rinfret, N. (1990). Language learning efficiency, dreams and REM sleep. *Psychiatric Journal of the University of Ottawa*, 15 (2), 91-92.
- Dewan, E. M. (1970). The programing (P) hypothesis for REM sleep. *International Psychiatry Clinics*, 7 (2), 295-307.
- Dietrich, A. & Kanso, R. (2010). A Review of EEG, ERP, and Neuroimaging Studies of Creativity and Insight. *Psychological Bulletin*, 136 (5), 822-848. [10.1037/a0019749](https://doi.org/10.1037/a0019749)
- Dragoi, G. & Tonegawa, S. (2011). Preplay of future place cell sequences by hippocampal cellular assemblies. *Nature*, 469 (7330), 397-401. [10.1038/nature09633](https://doi.org/10.1038/nature09633)
- (2013). Distinct preplay of multiple novel spatial experiences in the rat. *Proceedings of the National Academy of Sciences of the U.S.A.*, 110 (22), 9100-9105. [10.1073/pnas.1306031110](https://doi.org/10.1073/pnas.1306031110)
- Dresler, M. (2011). Kreativität, Schlaf und Traum – Neurobiologische Zusammenhänge. In K. Hermann (Ed.) *Neuroästhetik* (pp. 32-44). Kassel.

- (2012). Sleep and creativity. Theoretical models and neural basis. In D. Barrett & P. McNamara (Eds.) *Encyclopedia of Sleep and Dreams Vol II.* Santa Barbara.
- Dresler, M., Koch, S., Wehrle, R., Spoormaker, V. I., Holsboer, F., Steiger, A., Sämann, P. G., Obrig, H. & Czisch, M. (2011). Dreamed movement elicits activation in the sensorimotor cortex. *Current Biology*, 21 (21), 1833-1837. [10.1016/j.cub.2011.09.029](https://doi.org/10.1016/j.cub.2011.09.029)
- Dresler, M., Wehrle, R., Spoormaker, V. I., Holsboer, F., Steiger, A., Koch, S., Obrig, H., Sämann, P. G. & Czisch, M. (2012). Neural correlates of dream lucidity obtained from contrasting lucid versus non-lucid REM sleep: A combined EEG/fMRI case study. *Sleep*, 35 (7), 1017-1020. [10.5665/sleep.1974](https://doi.org/10.5665/sleep.1974)
- Dresler, M., Eibl, L., Fischer, C., Wehrle, R., Spoormaker, V. I., Steiger, A., Czisch, M. & Pawlowski, M. (2013). Volitional components of consciousness during wakefulness, dreaming and lucid dreaming. *Frontiers in Psychology*, 4, 987-987. [10.3389/fpsyg.2013.00987](https://doi.org/10.3389/fpsyg.2013.00987)
- Dresler, M., Spoormaker, V. I., Beiting, P. A., Czisch, M., Kimura, M., Steiger, A. & Holsboer, F. (2014). Neuroscience-driven discovery and development of sleep therapeutics. *Pharmacology & Therapeutics*, 141 (3), 300-334. [10.1016/j.pharmthera.2013.10.012](https://doi.org/10.1016/j.pharmthera.2013.10.012)
- Dresler, M., Erlacher, D., Czisch, M. & Spoormaker, V. I. (2015). Lucid dreaming. In M. Kryger, T. Roth & W. Dement (Eds.) *Principles and Practice of Sleep Medicine*. Amsterdam, NL: Elsevier.
- Dresler, M., Wehrle, R., Spoormaker, V. I., Holsboer, F., Steiger, A., Czisch, M. & Hobson, J. A. (2015a). Neural correlates of insight in dreaming and psychosis. *Sleep Medicine Reviews* 20, 92-99.
- Dresler, M. & Konrad, B. N. (2013). Mnemonic expertise during wakefulness and sleep. *Behavioral and Brain Sciences*, 36 (6), 616-617. [10.1017/S0140525X13001301](https://doi.org/10.1017/S0140525X13001301)
- Driskell, J. E., Copper, C. & Moran, A. (1994). Does mental practice enhance performance? *Journal of Applied Psychology*, 79 (4), 481-492. [10.1037/0021-9010.79.4.481](https://doi.org/10.1037/0021-9010.79.4.481)
- Dunbar, R. I. (1992). Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22 (6), 469-493. [10.1016/0047-2484\(92\)90081-J](https://doi.org/10.1016/0047-2484(92)90081-J)
- Dunbar, R. I. & Shultz, S. (2007). Evolution in the social brain. *Science*, 317 (5843), 1344-1347. [10.1126/science.1145463](https://doi.org/10.1126/science.1145463)
- Erlacher, D. & Chapin, H. (2010). Lucid dreaming: Neural virtual reality as a mechanism for performance enhancement. *International Journal of Dream Research*, 3 (1), 7-10. [10.11588/ijodr.2010.1.588](https://doi.org/10.11588/ijodr.2010.1.588)
- Erlacher, D., Stumbrys, T. & Schredl, M. (2011). Frequency of lucid dreams and lucid dream practice in German athletes. *Imagination, Cognition and Personality*, 31, 237-246. [10.2190/IC.31.3.f](https://doi.org/10.2190/IC.31.3.f)
- Erlacher, D. & Schredl, M. (2010). Practicing a motor task in a lucid dream enhances subsequent performance: A pilot study. *The Sport Psychologist*, 24 (2), 157-167.
- Fischer, S. & Born, J. (2009). Anticipated reward enhances offline learning during sleep. *Journal of Experimental Psychology. Learning, Memory and Cognition*, 35 (6), 1586-1593. [10.1037/a0017256](https://doi.org/10.1037/a0017256)
- Fischer, S., Hallschmid, M., Elsner, A. L. & Born, J. (2002). Sleep forms memory for finger skills. *Proceedings of the National Academy of Sciences of the U.S.A.*, 99 (18), 11987-11991. [10.1073/pnas.182178199](https://doi.org/10.1073/pnas.182178199)
- Flinn, M. V. & Alexander, R. D. (2007). Runaway social selection in human evolution. *The Evolution of Mind* (pp. 249-255). New York, NY: Guilford Press.
- Fosse, M. J., Fosse, R., Hobson, J. A. & Stickgold, R. J. (2003). Dreaming and episodic memory: A functional dissociation? *Journal of Cognitive Neuroscience*, 15 (1), 1-9. [10.1162/089892903321107774](https://doi.org/10.1162/089892903321107774)
- Frank, M. G. (2006). The mystery of sleep function: Current perspectives and future directions. *Reviews in the Neurosciences*, 17 (1), 375-92. [10.1515/revneuro.2006.17.4.375](https://doi.org/10.1515/revneuro.2006.17.4.375)
- Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11, 127-138. [10.1038/nrn2787](https://doi.org/10.1038/nrn2787)
- Genzel, L., Kroes, M. C., Dresler, M. & Battaglia, F. P. (2014). Light sleep versus slow wave sleep in memory consolidation: A question of global versus local processes? *Trends in Neurosciences*, 37 (1), 10-19. [10.1016/j.tins.2013.10.002](https://doi.org/10.1016/j.tins.2013.10.002)
- Gould, S. J. & Vrba, E. S. (1982). Exaptation – a missing term in the science of form. *Paleobiology*, 8 (1), 4-15.
- Groch, S., Wilhelm, I., Diekelmann, S. & Born, J. (2013). The role of REM sleep in the processing of emotional memories: Evidence from behavior and event-related potentials. *Neurobiology of Learning and Memory*, 99, 1-9. [10.1016/j.nlm.2012.10.006](https://doi.org/10.1016/j.nlm.2012.10.006)
- Hall, S. C. & van de Castle, R. I. (1966). *The content analysis of dreams*. New York, NY: Appleton-Century-Crofts.
- Heilman, K. M., Nadeau, S. E. & Beversdorf, D. O. (2003). Creative innovation: Possible brain mechanisms. *Neurocase*, 9 (5), 369-379. [10.1076/neur.9.5.369.16553](https://doi.org/10.1076/neur.9.5.369.16553)
- Hobson, J. A. (2009). REM sleep and dreaming: Towards a theory of protoconsciousness. *Nature Reviews Neuroscience*, 10 (11), 803-813. [10.1038/nrn2716](https://doi.org/10.1038/nrn2716)

- Hobson, J. A., Hong, C. C. & Friston, K. J. (2014). Virtual reality and consciousness inference in dreaming. *Frontiers in Psychology*, 5, 1133-1133. [10.3389/fpsyg.2014.01133](#)
- Hobson, J. A. & Friston, K. J. (2012). Waking and dreaming consciousness: Neurobiological and functional considerations. *Progress in Neurobiology*, 98 (1), 82-98. [10.1016/j.pneurobio.2012.05.003](#)
- Hobson, J. A. & Pace-Schott, E. F. (2002). The cognitive neuroscience of sleep: Neuronal systems, consciousness and learning. *Nature Reviews Neuroscience*, 3 (9), 679-693. [10.1038/nrn915](#)
- Hobson, J. A. & Wohl, H. (2005). *From angels to neurons. Art and the new science of dreaming*. Fidenza, I: Mattioli.
- Holzinger, B. (2014). Lucid dreaming in Psychotherapy. In R. Hurd & K. Bulkeley (Eds.) *Lucid Dreaming: New Perspectives on Consciousness in Sleep* (pp. 37-62).
- Horne, J. A. & McGrath, M. J. (1984). The consolidation hypothesis for REM sleep function: Stress and other confounding factors--a review. *Biological Psychology*, 18 (3), 165-184. [10.1016/0301-0511\(84\)90001-2](#)
- Hu, P., Stylos-Allan, M. & Walker, M. P. (2006). Sleep facilitates consolidation of emotional declarative memory. *Psychological Science*, 17 (10), 891-898. [10.1111/j.1467-9280.2006.01799.x](#)
- Indursky, P. & Rotenberg, V. (1998). Change of mood during sleep and REM sleep variables. *International Journal of Psychiatry in Clinical Practice*, 2 (1), 47-51. [10.3109/13651509809115114](#)
- Jouvet, M. (1979). What does a cat dream about? *Trends in Neurosciences*, 2, 280-282. [10.1016/0166-2236\(79\)90110-3](#)
- (1998). Paradoxical sleep as a programming system. *Journal of Sleep Research*, 7 (Suppl 1), 1-5. [10.1046/j.1365-2869.7.s1.1.x](#)
- Kahn, D., Stickgold, R., Pace-Schott, E. F. & Hobson, J. A. (2000). Dreaming and waking consciousness: A character recognition study. *Journal of Sleep Research*, 9 (4), 317-325. [10.1046/j.1365-2869.2000.00213.x](#)
- Kahn, D., Combs, A. & Krippner, S. (2002a). Dreaming as a function of chaos-like stochastic processes in the self-organizing brain. *Nonlinear Dynamics, Psychology, and Life Sciences*, 6 (4), 311-322. [10.1023/A:1019758527338](#)
- Kahn, D., Pace-Schott, E. & Hobson, J. A. (2002b). Emotion and cognition: Feeling and character identification in dreaming. *Consciousness and Cognition*, 11 (1), 34-50. [10.1006/ccog.2001.0537](#)
- Kahn, D. & Hobson, A. (2003). State dependence of character perception. *Journal of Consciousness Studies*, 10 (3), 57-68.
- (2005). Theory of mind in dreaming: Awareness of feelings and thoughts of others in dreams. *Dreaming*, 15 (1), 48-57. [10.1037/1053-0797.15.1.48](#)
- Karni, A., Tanne, D., Rubenstein, B. S., Askenasy, J. J. & Sagi, D. (1994). Dependence on REM sleep of overnight improvement of a perceptual skill. *Science*, 265 (5172), 679-682. [10.1126/science.8036518](#)
- Kris, E. (1952). *Psychoanalytic explorations in art*. New York, NY: International Universities Press.
- Lewin, I. (1989). The effect of 'waking-dreaming' on creativity and rote memory. *Cognition and Perception*, 9 (3), 225-236.
- Lima, S. L., Rattenborg, N. C., Lesku, J. A. & Amlaner, J. C. (2005). Sleeping under the risk of predation. *Animal Behaviour*, 70 (4), 723-736. [10.1016/j.anbehav.2005.01.008](#)
- Llewellyn, S. (2013). Such stuff as dreams are made on? Elaborative encoding, the ancient art of memory, and the hippocampus. *Behavioral and Brain Sciences*, 36 (6), 589-607. [10.1017/S0140525X12003135](#)
- Louie, K. & Wilson, M. A. (2001). Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron*, 29 (1), 145-156. [10.1016/S0896-6273\(01\)00186-6](#)
- Lyon, B. E. & Montgomerie, R. (2012). Sexual selection is a form of social selection. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 367 (1600), 2266-2273. [10.1098/rstb.2012.0012](#)
- Mahner, M. & Bunge, M. (2000). Function and functionalism: A synthetic perspective. *Philosophy of Science*, 68 (1), 75-94.
- Malcolm-Smith, S., Solms, M., Turnbull, O. & Tredoux, C. (2008). Threat in dreams: An adaptation? *Consciousness and Cognition*, 17 (4), 1281-1291. [10.1016/j.concog.2007.07.002](#)
- Malcolm-Smith, S., Koopowitz, S., Pantelis, E. & Solms, M. (2012). Approach/avoidance in dreams. *Consciousness and Cognition*, 21 (1), 408-412. [10.1016/j.concog.2011.11.004](#)
- Malinowski, J. E. & Horton, C. L. (2014). Memory sources of dreams: The incorporation of autobiographical rather than episodic experiences. *Journal of Sleep Research*, 23 (4), 441-447. [10.1111/jsr.12134](#)
- Maquet, P., Laureys, S., Peigneux, P., Fuchs, S., Petiau, C., Phillips, C., Aerts, J., Del Fiore, G., Degueldre, C., Meulemans, T., Luxen, A., Franck, G., van der Linden, M., Smith, C. & Cleeremans, A. (2000). Experience-dependent changes in cerebral activation during human REM sleep. *Nature Neuroscience*, 3 (8), 831-836. [10.1038/77744](#)

- Marks, G. A., Shaffery, J. P., Oksenberg, A., Speciale, S. G. & Roffwarg, H. P. (1995). A functional role for REM sleep in brain maturation. *Behavioural Brain Research*, 69 (1-2), 1-11. [10.1016/0166-4328\(95\)00018-O](https://doi.org/10.1016/0166-4328(95)00018-O)
- Martindale, C. (1999). Biological bases of creativity. *Handbook of Creativity* (pp. 137-152). Cambridge, UK: Cambridge University Press.
- McCarley, R. W. & Hoffman, E. (1981). REM sleep dreams and the activation-synthesis hypothesis. *American Journal of Psychiatry*, 138 (7), 904-912.
- Mednick, S. A. (1962). The associative basis of the creative process. *Psychological Review*, 69 (3), 220-232. [10.1037/h0048850](https://doi.org/10.1037/h0048850)
- Mellou, E. (1994). Play theories: A contemporary review. *Early Child Development and Care*, 102 (1), 91-100. [10.1080/0300443941020107](https://doi.org/10.1080/0300443941020107)
- Mendelsohn, G. A. (1976). Associative and attentional processes in creative performance. *Journal of Personality*, 44 (2), 341-369. [10.1111/j.1467-6494.1976.tb00127.x](https://doi.org/10.1111/j.1467-6494.1976.tb00127.x)
- Metzinger, T. (2003). *Being no one: The self-model theory of subjectivity*. Cambridge, MA: MIT Press.
- (2013). The myth of cognitive agency: Subpersonal thinking as a cyclically recurring loss of mental autonomy. *Frontiers in Psychology*, 4, 931-931. [10.3389/fpsyg.2013.00931](https://doi.org/10.3389/fpsyg.2013.00931)
- Millikan, R. (1984). *Language, thought and other biological categories*. Cambridge, MA: MIT Press.
- Mirmiran, M. (1995). The function of fetal/neonatal rapid eye movement sleep. *Behavioral Brain Research*, 69 (1-2), 13-22. [10.1016/0166-4328\(95\)00019-P](https://doi.org/10.1016/0166-4328(95)00019-P)
- Morselli, L. L., Guyon, A. & Spiegel, K. (2012). Sleep and metabolic function. *Pflügers Archiv*, 463 (1), 139-160. [10.1007/s00424-011-1053-z](https://doi.org/10.1007/s00424-011-1053-z)
- Neander, K. (1991). Functions as selected effects. *Philosophy of Science*, 58 (2), 168-184.
- Nielsen, T. A. (2000). A review of mentation in REM and NREM sleep: “covert” REM sleep as a possible reconciliation of two opposing models. *Behavioral and Brain Sciences*, 23 (6), 851-866.
- Nielsen, T. A., Kuiken, D., Alain, G., Stenstrom, P. & Powell, R. A. (2004). Immediate and delayed incorporations of events into dreams: Further replication and implications for dream function. *Journal of Sleep Research*, 13 (4), 327-336.
- Nielsen, T. A. & Powell, R. A. (1989). The ‘dream-lag’ effect: A 6-day temporal delay in dream content incorporation. *Psychiatry Journal of the University of Ottawa*, 14 (4), 561-565.
- Nir, Y. & Tononi, G. (2010). Dreaming and the brain: From phenomenology to neurophysiology. *Trends in Cognitive Sciences*, 14 (2), 88-100. [10.1016/j.tics.2009.12.001](https://doi.org/10.1016/j.tics.2009.12.001)
- Nishida, M., Pearsall, J., Buckner, R. L. & Walker, M. P. (2009). REM sleep, prefrontal theta, and the consolidation of human emotional memory. *Cerebral Cortex*, 19 (5), 1158-1166.
- Peigneux, P., Laureys, S., Fuchs, S., Destrebecqz, A., Collette, F., Delbeuck, X., Phillips, C., Aerts, J., Del, F. iore, Degueldre, C., Luxen, A., Cleeremans, A. & Maquet, P. (2003). Learned material content and acquisition level modulate cerebral reactivation during posttraining rapid-eye-movements sleep. *NeuroImage*, 20 (1), 125-134.
- Pellegrini, A. D. & Bjorklund, D. F. (2004). The ontogeny and phylogeny of children’s object and fantasy play. *Human Nature*, 15 (1), 23-43. [10.1007/s12110-004-1002-z](https://doi.org/10.1007/s12110-004-1002-z)
- Pennartz, C. M., Lee, E., Verheul, J., Lipa, P., Barnes, C. A. & McNaughton, B. L. (2004). The ventral striatum in off-line processing: Ensemble reactivation during sleep and modulation by hippocampal ripples. *Journal of Neuroscience*, 24 (29), 6446-6456. [10.1523/JNEUROSCI.0575-04.2004](https://doi.org/10.1523/JNEUROSCI.0575-04.2004)
- Perogamvros, L. & Schwartz, S. (2012). The roles of the reward system in sleep and dreaming. *Neuroscience & Biobehavioral Reviews*, 36 (8), 1934-1951. [10.1016/j.neubiorev.2012.05.010](https://doi.org/10.1016/j.neubiorev.2012.05.010)
- (2014). Sleep and emotional functions. *Current Topics in Behavioral Neurosciences*. [10.1007/7854_2013_271](https://doi.org/10.1007/7854_2013_271)
- Plihal, W. & Born, J. (1997). Effects of early and late nocturnal sleep on declarative and procedural memory. *Journal of Cognitive Neuroscience*, 9 (4), 534-547. [10.1162/jocn.1997.9.4.534](https://doi.org/10.1162/jocn.1997.9.4.534)
- Rak, M., Beiting, P. A., Steiger, A., Schredl, M. & Dresler, M. (in press). Increased lucid dreaming frequency in narcolepsy. *Sleep*.
- Rasch, B. & Born, J. (2013). About sleep’s role in memory. *Physiological Review*, 93 (2), 681-766. [10.1152/physrev.00032.2012](https://doi.org/10.1152/physrev.00032.2012)
- Rasch, B., Büchel, C., Gais, S. & Born, J. (2007). Odor cues during slow-wave sleep prompt declarative memory consolidation. *Science*, 315 (5817), 1426-1429. [10.1126/science.1138581](https://doi.org/10.1126/science.1138581)
- Revonsuo, A. (1995). Consciousness, dreams and virtual realities. *Philosophical Psychology*, 8 (1), 35-58. [10.1080/09515089508573144](https://doi.org/10.1080/09515089508573144)

- (2000). The reinterpretation of dreams: An evolutionary hypothesis of the function of dreaming. *Behavioral and Brain Sciences*, 23 (6), 877-901.
- (2006). *Inner presence*. Boston, MA: MIT Press.
- Revonsuo, A., Tuominen, J. & Valli, K. (2015). The avatars in the machine: Dreaming as a simulation of social reality. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Ritter, S. M. & Dijksterhuis, A. (2014). Creativity-the unconscious foundations of the incubation period. *Frontiers in Human Neuroscience*, 8, 215-215. [10.3389/fnhum.2014.00215](https://doi.org/10.3389/fnhum.2014.00215)
- Ritter, S. M., Strick, M., Bos, M. W., van Baaren, R. B. & Dijksterhuis, A. (2012). Good morning creativity: Task reactivation during sleep enhances beneficial effect of sleep on creative performance. *Journal of Sleep Research*, 21 (6), 643-647. [10.1111/j.1365-2869.2012.01006.x](https://doi.org/10.1111/j.1365-2869.2012.01006.x)
- Roffwarg, H. P., Muzio, J. N. & Dement, W. C. (1966). Ontogenetic development of the human sleep-dream cycle. *Science*, 152 (3722), 604-619. [10.1126/science.152.3722.604](https://doi.org/10.1126/science.152.3722.604)
- Schredl, M. & Erlacher, D. (2011). Frequency of lucid dreaming in a representative German sample. *Perceptual and Motor Skills*, 112 (1), 104-108. [10.2466/09.PMS.112.1.104-108](https://doi.org/10.2466/09.PMS.112.1.104-108)
- Schredl, M., Hoffmann, L., Sommer, J. U. & Stuck, B. A. (2014). Olfactory stimulation during sleep can reactivate odor-associated images. *Chemosensory Perception*. [10.1007/s12078-014-9173-4](https://doi.org/10.1007/s12078-014-9173-4)
- Schredl, M. & Hofmann, F. (2003). Continuity between waking activities and dream activities. *Consciousness and Cognition*, 12 (2), 298-308. [10.1016/S1053-8100\(02\)00072-7](https://doi.org/10.1016/S1053-8100(02)00072-7)
- Schuster, C., Hilfiker, R., Amft, O., Scheidhauer, A., Andrews, B., Butler, J., Kischka, U. & Ettlin, T. (2011). Best practice for motor imagery: A systematic literature review on motor imagery training elements in five different disciplines. *BMC Med*, 9 (75). [10.1186/1741-7015-9-75](https://doi.org/10.1186/1741-7015-9-75)
- Schädlich, M. & Erlacher, D. (2012). Applications of lucid dreams: An online study. *International Journal of Dream Res*, 5 (2), 134-134. [10.11588/ijodr.2012.2.9505](https://doi.org/10.11588/ijodr.2012.2.9505)
- Sikka, P., Valli, K., Virta, T. & Revonsuo, A. (2014). I know how you felt last night, or do I? Self- and external ratings of emotions in REM sleep dreams. *Consciousness and Cognition*, 25, 51-66. [10.1016/j.concog.2014.01.011](https://doi.org/10.1016/j.concog.2014.01.011)
- Smith, C. T., Nixon, M. R. & Nader, R. S. (2004). Posttraining increases in REM sleep intensity implicate REM sleep in memory processing and provide a biological marker of learning potential. *Learning & Memory*, 11 (6), 714-719. [10.1101/lm.74904](https://doi.org/10.1101/lm.74904)
- Snyder, F. (1970). The phenomenology of dreaming. *The Psychodynamic Implications of The Physiological Studies on Dreams* (pp. 124-151). Springfield, IL: Charles C. Thomas.
- Spoormaker, V. I., van den Bout, J. & Meijer, E. J. G. (2003). Lucid dreaming treatment for nightmares: A series of cases. *Dreaming*, 13 (3), 181-186. [10.1023/A:1025325529560](https://doi.org/10.1023/A:1025325529560)
- Spoormaker, V. I. & van den Bout, J. (2006). Lucid dreaming treatment for nightmares: A pilot study. *Psychotherapy and Psychosomatics*, 75 (6), 389-394. [10.1159/000095446](https://doi.org/10.1159/000095446)
- Stumbrys, T. & Daniels, M. (2010). An exploratory study of creative problem solving in lucid dreams: Preliminary findings and methodological considerations. *International Journal of Dream Research*, 3 (2), 121-129. [10.11588/ijodr.2010.2.6167](https://doi.org/10.11588/ijodr.2010.2.6167)
- Stumbrys, T., Erlacher, D., Johnson, M. & Schredl, M. (2014). The phenomenology of lucid dreaming: An online survey. *American Journal of Psychology*, 127 (2), 191-204.
- Stumbrys, T., Erlacher, D. & Schredl, M. (2015). Effectiveness of motor practice in lucid dreams: A comparison with physical and mental practice. *Journal of Sports Sciences*. [10.1080/02640414.2015.1030342](https://doi.org/10.1080/02640414.2015.1030342)
- Tholey, P. (1989). Consciousness and abilities of dream characters observed during lucid dreaming. *Perceptual and Motor Skills*, 68 (2), 567-578. [10.2466/pms.1989.68.2.567](https://doi.org/10.2466/pms.1989.68.2.567)
- Tononi, G. & Cirelli, C. (2006). Sleep function and synaptic homeostasis. *Sleep Medicine Review*, 10 (1), 49-62.
- Valli, K. & Revonsuo, A. (2009). The threat simulation theory in light of recent empirical evidence: A review. *American Journal of Psychology*, 122 (1), 17-38.
- van der Werf, Y. D., Altena, E., Schoonheim, M. M., Sanz-Arigita, E. J., Vis, J. C., de Rijke, W. & van Someren, E. J. (2009). Sleep benefits subsequent hippocampal functioning. *Nature Neuroscience*, 12 (2), 122-123. [10.1038/nn.2253](https://doi.org/10.1038/nn.2253)
- Vassalli, A. & Dijk, D. J. (2009). Sleep function: Current questions and new approaches. *European Journal of Neuroscience*, 29 (9), 1830-1841. [10.1111/j.1460-9568.2009.06767.x](https://doi.org/10.1111/j.1460-9568.2009.06767.x)

- Voss, U., Holzmann, R., Tuin, I. & Hobson, J. A. (2009). Lucid dreaming: A state of consciousness with features of both waking and non-lucid dreaming. *Sleep*, 32 (9), 1191-1200.
- Voss, U., Holzmann, R., Hobson, A., Paulus, W., Koppehele-Gossel, J., Klimke, A. & Nitsche, M. A. (2014). Induction of self awareness in dreams through frontal low current stimulation of gamma activity. *Nature Neuroscience*, 17 (6), 810-812. [10.1038/nn.3719](#)
- Wagner, U., Gais, S. & Born, J. (2001). Emotional memory formation is enhanced across sleep intervals with high amounts of rapid eye movement sleep. *Learning and Memory*, 8 (2), 112-119. [10.1101/lm.36801](#)
- Wagner, U., Gais, S., Haider, H., Verleger, R. & Born, J. (2004). Sleep inspires insight. *Nature*, 427 (6972), 352-355-352-355. [10.1038/nature02223](#)
- Wagner, U., Hallschmid, M., Rasch, B. & Born, J. (2006). Brief sleep after learning keeps emotional memories alive for years. *Biological Psychiatry*, 60 (7), 788-790. [10.1016/j.biopsych.2006.03.061](#)
- Walker, M. P. & van der Helm, E. (2009). Overnight therapy? The role of sleep in emotional brain processing. *Psychological Bulletin*, 135 (5), 731-748. [10.1037/a0016570](#)
- Walsh, D. M. & Ariew, A. (1996). A taxonomy of functions. *Canadian Journal of Philosophy*, 26 (4), 493-514.
- Wamsley, E. J. (2014). Dreaming and offline memory consolidation. *Current Neurology and Neuroscience Reports*, 14, 433-433. [10.1007/s11910-013-0433-5](#)
- Wamsley, E. J., Perry, K., Djonlagic, I., Reaven, L. B. & Stickgold, R. (2010a). Cognitive replay of visuomotor learning at sleep onset: Temporal dynamics and relationship to task performance. *Sleep*, 33 (1), 59-68.
- Wamsley, E. J., Tucker, M., Payne, J. D., Benavides, J. A. & Stickgold, R. (2010b). Dreaming of a learning task is associated with enhanced sleep-dependent memory consolidation. *Current Biology*, 20 (9), 850-855. [10.1016/j.cub.2010.03.027](#)
- Wamsley, E. J. & Stickgold, R. (2011). Memory, sleep and dreaming: Experiencing consolidation. *Sleep Medicine Clinics*, 6 (1), 97-108.
- West-Eberhard, M. J. (2014). Darwin's forgotten idea: The social essence of sexual selection. *Neuroscience and Biobehavioral Reviews* 46, 501-508. [10.1016/j.neubiorev.2014.06.015](#)
- Wilson, M. A. & McNaughton, B. L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science*, 265 (5172), 676-679. [10.1126/science.8036517](#)
- Windt, J. M. (2010). The immersive spatiotemporal hallucination model of dreaming. *Phenomenology and the Cognitive Sciences*, 9, 295-316. [10.1007/s11097-010-9163-1](#)
- Windt, J. M., Harkness, D. L. & Lenggenhager, B. (2014). Tickle me, I think I might be dreaming! Sensory attenuation, self-other distinction, and predictive processing in lucid dreams. *Frontiers in Human Neuroscience*, 8, 717-717. [10.3389/fnhum.2014.00717](#)
- Wolpin, M., Marston, A., Randolph, C. & Clothier, A. (1992). Individual difference correlates of reported lucid dreaming frequency and control. *Journal of Mental Imagery*, 16, 231-236.
- Wouters, A. G. (2003). Four notions of biological function. *Studies in History and Philosophy of Biology and Biomedical Sciences*, 34 (4), 633-668. [10.1016/j.shpsc.2003.09.006](#)
- (2013). Function, Biological. In W. Dubitzki, O. Wolkenhauer, H. Yokota & K.-H. Cho (Eds.) *Encyclopedia of Systems Biology*. Berlin, GER: Springer.
- Xie, L., Kang, H., Xu, Q., Chen, M. J., Liao, Y., Thiyagarajan, M., O'Donnell, J., Christensen, D. J., Nicholson, C., Iliff, J. J., Takano, T., Deane, R. & Nedergaard, M. (2013). Sleep drives metabolite clearance from the adult brain. *Science*, 342 (6156), 373-377. [10.1126/science.1241224](#)
- Zadra, A., Desjardins, S. & Marcotte, E. (2006). Evolutionary function of dreams: A test of the threat simulation theory in recurrent dreams. *Consciousness and Cognition*, 15 (2), 450-463.

The Simulation Theories of Dreaming: How to Make Theoretical Progress in Dream Science

A Reply to Martin Dresler

Antti Revonsuo, Jarno Tuominen & Katja Valli

Among the most pressing challenges for dream science is the difficulty of establishing theoretical unification between the various theories, ideas, and findings that have been presented in the literature to answer the question of how it is possible to construct a solid scientific theory with predictive and explanatory power in dream science. We suggest that the concept of “world-simulation” serves as the core concept for a theoretically unified paradigm to describe and explain dreaming. From this general concept, more specific theories of the function of dreaming can be derived, such as the Threat Simulation Theory (TST) and the Social Simulation Theory (SST), as we argued in our target article. We agree with Dresler that these two functions may not be the only functions of dreaming, but we still have grounds to believe that they are the strongest contenders. In our reply we first clarify why the functions of sleep should be considered separately from the functions of dreaming. Second, we outline what a good scientific theory of dreaming should be like and what it should be capable of. Furthermore, we evaluate the current state of simulation theories within this context. To conclude, we propose that instead of a general multifunctional theory of sleep and dreaming, where no hypothesis is excluded, the future progress of dream science will benefit more from opposing, competing and mutually exclusive theories about the specific functions of dreaming. This, however, demands that the opposing theories and their predictions must be risky, clearly formulated, and empirically testable.

Keywords

Avatars | Dream | Dreaming | Multifunctionality | Simulation | Sleep | Social simulation | Threat simulation | Virtual reality

Authors

Antti Revonsuo
antti.revonsuo@utu.fi
Högskolan i Skövde
Skövde, Sweden
Turun yliopisto
Turku, Finland

Jarno Tuominen
jarno.tuominen@utu.fi
Turun yliopisto
Turku, Finland

Katja Valli
katval@utu.fi
Turun yliopisto
Turku, Finland
Högskolan i Skövde
Skövde, Sweden

Commentator

Martin Dresler
martin.dresler@donders.ru.nl
Radboud Universiteit Medical Center
Nijmegen, Netherlands

Editors

Thomas Metzinger
metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Introduction

We are grateful to [Martin Dresler \(this collection\)](#) for his thorough and insightful commentary on our target article ([Revonsuo et al. this collection](#)). Dresler's commentary places the proposed simulation functions of dreaming into the wider context of other functions for sleep and dreaming, demonstrating that these phenomena may have multiple different and partly overlapping functions. He also suggests the threat simulation and social simulation functions are unique. They can neatly be connected to evolutionary theory and only they explain why the suppression of reality testing and the lack of lucidity are necessary features of these simulation functions of dreaming (i.e., they require an "oblivious avatar"). While we agree with many of the points presented in Dresler's analysis, we believe that it is possible to regard the different proposed functions of dreaming as representing different (preliminary) scientific theories of dreaming. When viewed from this theory-driven perspective, it is also possible to present more definitive evaluations as to which of them are more plausible theoretical explanations than others.

2 Function of sleep vs. function of dreaming

Many of the findings [Dresler \(this collection\)](#) mentions in his commentary are not about dreaming, but rather about sleep, its different stages, and their potential correlates, effects, and functions. While it is encouraging that there is much evidence about the functions of sleep that relates to memory and learning, and that emotionally significant information seems to hold a special place, most of those studies have very little or nothing to do with dreaming as a subjective experience. In most of the sleep studies, whether or not the sleeping participants have been dreaming or not, and what their dream contents have been, is irrelevant for the hypotheses being tested (e.g., whether a certain stage of sleep enhances memory consolidation of particular types of stimuli) and usually remains unknown. In sleep studies purely objective

neurophysiological and behavioural phenomena are investigated with objective measures. In contrast, in dream studies purely subjective phenomena are explored by collecting subjective introspective reports describing the contents of phenomenal consciousness. Modern theories of the functions of sleep are undoubtedly quite strong as scientific theories of sleep and its relationship to some neurocognitive mechanisms of memory and learning, but they are not in any direct sense theories of dreaming. Of course, any proposed theory of dreaming should be at the very least *consistent* with the leading theories of sleep, because the phenomenal level of organization supervenes on the lower, neurophysiological level. However, the opposite is not necessarily true. As [Dresler \(this collection\)](#) points out, lower-level functions can be carried out independently of the higher, phenomenal level of organization. Thus, we would like to strongly emphasize that the merits and the predictions of theories of dreaming primarily have to be tested by using data that reflects subjective dream contents, not the objective features of sleep.

3 What is it like to be a strong scientific theory of dreaming?

Any theory of a phenomenon should include a precise definition and description of its target phenomenon (or *explanandum*), as well as clear demarcation of conceptually and empirically different phenomena. Theories of dreaming should clearly state i) in what way dreaming is a different type of phenomenon from sleep (or any particular stage of sleep), and ii) in what way dreaming is a special form of mental activity occurring during sleep. In our approach the starting points are that while sleep and its different stages can be defined by objective behavioural and neurophysiological criteria, dreaming is a subjective phenomenon; a special, complex altered state of consciousness that can be differentiated from simple sleep mentation. Quite independently from any functional considerations, the general, universal *form* of dreaming, as most dream researchers currently agree, is a complex,

multi-modal *simulation* of the sensory perceptual world, inhabited by a simulated self or a self-model (Hobson 2009; Metzinger 2003, 2013; Nielsen 2010; Windt 2010). A fruitful idea in biology is that *form suggests function*; thus the form that dreaming takes, a world-simulation, most likely suggests that the major functions of dreaming have something to do with world-simulation. The most frequent dream contents are therefore the most likely candidates for reflecting the specific function(s) of dreaming: how, when, under what circumstances, and what contents to simulate. Thus, to state that dreaming is an internal world-simulation is to describe the general form that this phenomenon universally takes, but not necessarily its function. The function(s) of the simulation, according to our view, are mainly related to the specific contents selected for simulation.¹

Furthermore, a proper theory of dreaming should be *simple yet covering*, so that the same general principles apply to many types of dreams, including the pathologies of dreaming, animal dreaming, and other special cases; the theory should be *fruitful*, so that it leads to new ideas, hypotheses, and new directions for active research; it should be *empirically testable*, so that it leads to risky predictions whose accuracy can be objectively checked. It should have both predictive and explanatory power.

Of course, these virtues are desirable in any scientific theory of any phenomenon. When there are rival theories of the same phenomenon, they should be compared with regard to their overall strengths and weaknesses as scientific theories. If they are consistent with each other, perhaps they can be combined into a single, more covering theory. If they are inconsistent with each other, their differing predictions should be empirically tested. After their relative strengths and weaknesses are compared, it should be possible to say which ones are stronger than others.

¹ Further, Dresler (this collection) raises the question of whether the frequency of specific dream contents can be regarded as evidence for the importance of its underlying functions. If we consider the function of dreaming more broadly to be that of a training ground for essential and adaptive behaviors, it becomes rather clear that the observed frequency of these behaviors can be viewed as a valid measure of their importance. This, however, is evident only when comparing the contents within the phenomenal level of explanation.

4 Simulation theories of dreaming

According to the simulation view, dreaming is a special case of phenomenal consciousness, or the phenomenal level of organization being activated in the brain. Waking consciousness and dreaming are manifestations of the same natural biological phenomenon in the brain, but they occur in different contexts and under different conditions. The simulation theory of dreaming is anchored to a more general theory of consciousness, which in turn is anchored philosophically to weak emergent materialism and multi-level explanation (Bechtel 2008, 2011; Craver 2007; Revonsuo 2006, 2010). In a multi-level explanation of a mental phenomenon, several different explanatory dimensions surround the target phenomenon: the *downward-looking* explanation specifies its neural correlates and mechanisms; the *backward-looking* mechanisms specify what has causally brought about or modulated the phenomenon (e.g., day residues or traumatic experiences that directly influenced specific contents of dreaming; the ontogeny of dreaming—how dreaming came about during individual development; phylogeny—how dreaming emerged and might have been selected for during evolutionary history²); and the *upward-looking* (functional) explanation—how does dreaming guide or change consequent mental states or external behaviours? Only after all these explanatory dimensions can be accounted for may we be said to have a comprehensive theory of dreaming, including its function(s) (see also Revonsuo 2006, 2010; Valli 2011).

So far, one general and three separate, more specific simulation theories have been proposed. From a more general perspective, some versions of the Continuity Hypothesis (CH) can be regarded as a simulation theory, as some proponents of it consider the world-simulation itself to be a functional *form* of dreaming (e.g., Foulkes 1985, pp. 201–202). Three other, more specific simulation theories have been proposed:

² We should, however, also keep in mind the option that dreaming does not serve any function at all and was not selected for, but is merely epiphenomenal, as suggested, for example, by Flanagan (2001), and implied by the Continuity Hypothesis (CH). This notion should be the null hypothesis against which the proposed functions of dreaming are to be pitted.

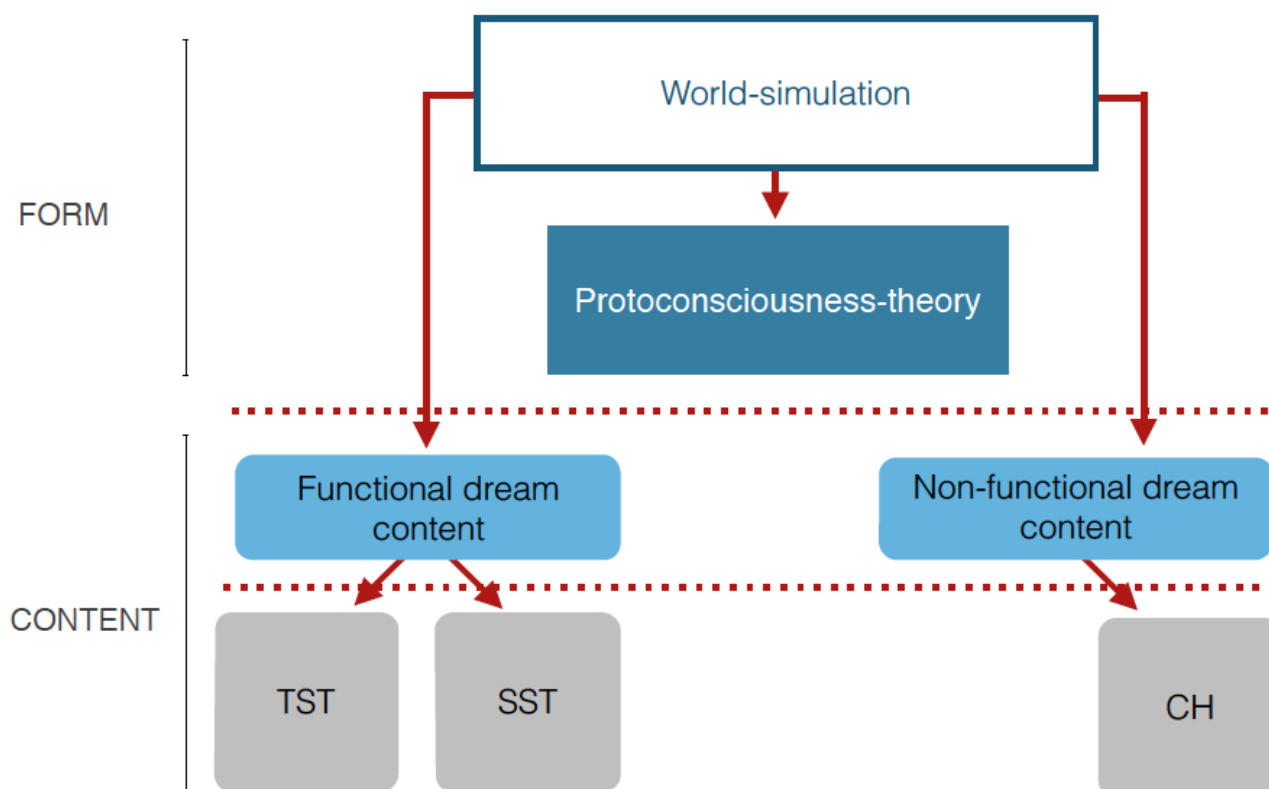


Figure 1: Simulation theories of dreaming. All simulation theories assume that dreaming can be defined as a world-simulation, the form of which is functional. The protoconsciousness-theory is more focused on explaining the form of dreams instead of their specific contents. Threat simulation and Social simulation theories try to explain the content of dreams as having a specific function, while the Continuity Hypothesis assumes the content of the simulation to be evolutionarily non-functional.

the protoconsciousness theory (Hobson 2009), which covers the role of dreaming in ontogeny; the Threat-Simulation Theory (TST), which covers the negative contents of dreaming and provides an evolutionary account for them; and the Social Simulation Theory (SST), which covers the social contents of dreaming, including the positively charged ones. Taken together, these theories are at the same time both covering and economical: the simple principle of “internally activated world-simulation” underlies all of them (see figure 1). The proto-consciousness theory accounts for how and why the basic form of the virtual-reality generator comes about in the developing brain, and how during early brain maturation both dreaming and waking consciousness emerge together in interaction. It is, however, the most speculative of the three simulation theories, as we cannot hope to

test it with data about subjective experiences describing the postulated fetal dream experience: what is it like to be a proto-conscious dreaming fetus? Thus, its weakness is that dream reports or any other direct evidence of the existence of subjective dream-like states cannot conceivably be empirically collected to test the validity of the theory.

The TST and SST, as we have explicated in our target article (Revonsuo et al. this collection) and in earlier publications elsewhere (Revonsuo 2000, 2006; Valli & Revonsuo 2009) are testable as they issue specific predictions concerning the frequency and quality of dream contents under different circumstances. They are also covering, TST potentially accounts for normal dreaming as well as several special types of dreams, where negative dream contents are particularly abundant and dominate (bad

dreams, recurrent dreams, nightmares, post-traumatic dreams, children's earliest dreams, dreams in parasomnias such as RBD, night terrors, and so on). Together TST and SST potentially cover a very large proportion of the statistically most frequent dream contents, and the predictions derived from these theories have specific empirically testable consequences as to the quantity and quality of these types of dream contents. As [Dresler \(this collection\)](#) points out, simulation theories also have the advantage of being highly consistent with the peculiar behavioural, neurophysiological, and phenomenal features of dreaming such as isolation from sensory input, motor activity, cognitive reflection, and reality testing. These features are necessary preconditions for running powerful, phenomenologically realistic but behaviorally isolated virtual reality simulations in the sleeping brain. The simulation theories thus have a lot of explanatory power. The concept of world-simulation unifies numerous separate phenomena related to dreaming and makes sense of them under a single concept. In this the simulation theories of dreaming fulfil the requirements of simplicity, coverage, and economy as well as having predictive and explanatory power. Compared to some of the other ideas Dresler presents in his commentary, it appears that currently the simulation theories are amongst the strongest frameworks for the form and function of dreaming.

5 Rival paradigms in dream science

Of the theories that are directly applicable to dreaming, we have already addressed the Continuity Hypothesis (CH) in our target article ([Revonsuo et al. this collection](#)). As we say there, it has never been formulated in a sufficiently precise manner such that risky, testable predictions can be derived from it. The CH, largely because of its vagueness, might actually be consistent with simulation theories. The particular contents of dreams are neither selected through an active process, nor do they reflect any function(s); they are selected through a passive and more or less random mirroring of the experiences that have been lived through. Further, CH does not consider how to deal with

potential anomalies for the theory: the relatively frequent cases where either something very alien to our waking world (and thus entirely discontinuous with it) appears, or where something very common in our waking life fails to appear in our dream contents. Can the theory be regarded as falsified when evidence of such blatantly discontinuous dream contents appear over and over again in dream data? One version of the CH, presented by [Foulkes \(1985\)](#) states that the mnemonic sources of dream contents are random and unpredictable; thus dream contents are unselective random samples of our memories; but the general form of dreams as world simulations as such is highly predictable—thus the function of dreaming would be more related to the general form than to the specific contents of dreams. However, as we have argued, dream contents are *not* random, but selective, and in particular they select threatening and social events into dreams. Thus, the basic assumption behind Foulkes's version of CH has turned out to be empirically false. The CH thus does not look very promising. But, as we argued in our target article ([Revonsuo et al. this collection](#)), some testable predictions can and should be derived from CH, to render its predictions as the null hypothesis “no selectivity, no functionality”, and thereby directly test its predictions against those derived from TST and SST.

Another major functional theory of dreaming, the Emotion Regulation Theory (ERT; also reviewed by [Dresler this collection](#)), also seems relatively weak as a scientific theory. It has been presented by many different authors in many different formulations (e.g., [Cartwright et al. 2006](#); [Hartmann 1996](#); [Kramer 1991](#)). There seems to be no standard, detailed, or shared version of this theory among its supporters; thus it also suffers from a vagueness similar to that of CH. The shared core in all of the different versions appears to be the idea that dreaming works with and processes difficult, unpleasant emotions and events, and through this dream processing makes us get over them and feel and function better in our lives. An often-used analogy compares dreaming with psychotherapy ([Hartmann 1995](#); [Walker & van der Helm 2009](#)).

Again, when looking at the evidence it is important to separate sleep from dreaming. When it comes to emotional processing during sleep, the analogy to psychotherapy gains some support (Walker & van der Helm 2009). But when applied specifically to dreaming and dream contents, the idea runs into difficulties. Its theoretical roots appear to originate predominantly from the clinical tradition, and more specifically from the idea that the function of dreaming is to protect sleep from strong surges of emotion and to solve emotional problems. The negative contents of dreams originate from interpersonal conflicts and current concerns, thus being consistent with the continuity between dreaming and waking, in fact so much so that the CH coupled with the ERT could perhaps be seen to form one specific paradigm of dream theorizing. Perhaps one of the core differences between the ERT+CH paradigm and the simulation paradigm is their relationship to biological explanations. The ERT+CH favours psychological-level explanations and emphasizes recent individual experiences (learning, nurture) as proximate explanations of dreaming. The simulation paradigm emphasizes biological explanations of the form and contents of dreaming, and links dream consciousness to both the underlying neurophysiological levels as well as the ontogenetic and phylogenetic, ultimate biological history of dreaming as explanations of the form and contents of dreaming. A further core difference between these paradigms is that the psychological paradigm sees the function of dreaming as contributing to our psychological well-being and psychological adaptation to our lives, whereas the biological paradigm sees the origin of dreaming in its ability to increase fitness in all mammals and in humans during their evolutionary history; but dreaming need not *necessarily* contribute to our psychological well-being in order to fulfill its original biological function.

As these approaches represent different paradigms with differing core ideas, it might not be possible to integrate them, in the manner that Dresler (this collection) suggests, into one overall multifunctional theory of dreaming. Some of the core assumptions of ERT are incon-

sistent with TST, especially when it comes to the function(s) of dreaming and to the explanation of nightmares and bad dreams. According to TST, post-traumatic dreams, recurrent dreams, nightmares, bad dreams, and the earliest dreams in childhood are the best and strongest manifestations of the function of dreaming, when the function is fully at work and typically activated by ecologically valid threat cues and dangerous events observed in the environment, often displaying universal threat scripts consistent with evolutionarily relevant threats. In parasomnias the threat-simulation system can be overactivated, or activated in an inappropriate context and therefore seen as psychologically dysfunctional, so that it might in actuality either decrease the well-being of the individual or hamper with other functions of sleep and dreaming, even though it at the same time carries out its original biological function perfectly. By contrast, according to ERT, such highly negative dreams are malfunctions and failures of the core function of dreaming itself, because such dreams disturb sleep and make us feel negative emotions. Nightmares cause psychological suffering and sleep disturbances, thus they are like a failed psychotherapy session that increases the individual's psychological distress, instead of calming the individual down. As such, very large and important categories of dreams (and their functionality) are explained in squarely opposing ways by the two paradigms.

6 Concluding remarks

Consequently, it is not only possible but theoretically necessary to separate the basic assumptions, the predictions, and the hypotheses of the simulation theories from those of ERT and others. We can have multiple *theories* of dream functions, but dreaming as a specific phenomenon cannot have multiple *conflicting functions*! If one theory says that recurrent dreams, nightmares, and bad dreams are types of dreams that most strongly carry out the TST functions and thus were selected for in human evolutionary history, and another theory says that such dreams are, from the functional point

of view, total failures of dream function, it becomes impossible to construct from those mutually opposing ingredients a “multifunctional” theory.³ A theory that combines TST and ERT would have to say that on the one hand the function of dreaming is to have many threatening events in dreams, bad dreams, nightmares, and recurrent negative dreams, in order to rehearse threat perception and avoidance, but on the other hand the function of dreaming is also to calm down or suppress exactly those types of dreams to make the dreamer feel better. What is the dream production system supposed to do: increase or decrease the number and impact of these kinds of dreams? The multifunctional theory cannot derive coherent testable predictions about the quantity and quality of these types of dreams.

This situation, however, is far from a scientific catastrophe; in fact, it is highly *desirable*. The problem is not that there is a lack of different theories, hypotheses, ideas, or suggestions about the nature and functions of dreaming, but rather that there are too many. Consequently, it is not only possible, but theoretically necessary to separate the basic assumptions, predictions, and hypotheses of the simulation theories from those of ERT, CH, and others. We can have multiple independent *theories* of dream functions, but dreaming as a specific phenomenon cannot have multiple *mutually inconsistent functions*. We hope that the simulation theories of dreaming, whether they turn out to be correct or not, will at least push dream science forward. The progress of any science is best served by the directly opposing predictions issued by rival, clearly stated, empirically testable hypotheses. Thus it is, from the scientific point of view, much more desirable to

have many squarely opposing testable hypotheses than one all-inclusive theory that is unfalsifiable or too vague to be tested. When the opposing theories have been well-formulated and put through fair but strict empirical tests several times, we will know which ones to adopt for the time being and which ones to leave behind for good, in order to keep dream science a progressive branch of science.

³ The multifunctionality of dreaming might be possible in different populations, so that in a population that lives in a very threat-filled environment a strong threat simulation system would be selected for, whereas in a population living in more peaceful conditions the psychotherapeutic function and taming of threat simulations dreams would be more likely candidates for selection. However, one and the same population cannot manifest both functions at the same time. Just as in some species of moths, in one environment individuals are selected for towards being white because white provides the best camouflage, while in another environment the color of individuals in the same moth species is selected for towards being dark gray or black, because in that environment all the white individuals are too easily detected by predators.

References

- Bechtel, W. (2008). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. London, UK: Routledge University Press.
- (2011). Mechanism and biological explanation. *Philosophy of Science*, 78 (4), 533-557.
- Cartwright, R., Agargun, M., Kirkby, J. & Friedman, J. (2006). Relation of dreams to waking concerns. *Psychiatry Research*, 141 (3), 261-270. [10.1016/j.psychres.2005.05.013](https://doi.org/10.1016/j.psychres.2005.05.013)
- Craver, C. F. (2007). *Explaining the brain: What a science of the mind-brain could be*. New York, NY: Oxford University Press.
- Dresler, M. (2015). The multifunctionality of dreaming and the oblivious avatar-A commentary on Antti Revonsuo and colleagues. *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Flanagan, O. (2001). *Dreaming souls: Sleep, dreams and the evolution of the conscious mind*. New York, NY: Oxford University Press.
- Foulkes, D. (1985). *Dreaming: A cognitive-psychological analysis*. Hillsdale, NJ: Lawrence Erlbaum.
- Hartmann, E. (1995). Making connections in a safe place: Is dreaming psychotherapy? *Dreaming*, 5 (4), 213-228. [10.1037/h0094437](https://doi.org/10.1037/h0094437)
- (1996). Outline for a theory on the nature and functions of dreaming. *Dreaming*, 6 (2), 147-170. [10.1037/h0094452](https://doi.org/10.1037/h0094452)
- Hobson, J. A. (2009). REM sleep and dreaming: Towards a theory of protoconsciousness. *Nature Reviews Neuroscience*, 10 (11), 803-813. [10.1038/nrn2716](https://doi.org/10.1038/nrn2716)
- Kramer, M. (1991). The nightmare: A failure in dream function. *Dreaming*, 1 (4), 277-285. [10.1037/h0094339](https://doi.org/10.1037/h0094339)
- Metzinger, T. (2003). *Being no one: The self-model theory of subjectivity*. Cambridge, MA: MIT Press.
- (2013). Why are dreams interesting to philosophers? The example of minimal phenomenal selfhood, plus an agenda for future research. *Frontiers in Psychology*, 4 (746). [10.3389/fpsyg.2013.00746](https://doi.org/10.3389/fpsyg.2013.00746)
- Nielsen, T. A. (2010). Dream analysis and classification: The reality simulation perspective. In M. Kryeger, T. Roth & W. C. Dement (Eds.) *Principles and practice of sleep medicine* (pp. 595-603). New York, NY: Elsevier.
- Revonsuo, A. (2000). The reinterpretation of dreams: An evolutionary hypothesis of the function of dreaming. *Behavioral and Brain Sciences*, 23 (6), 877-901. [10.1017/S0140525X00004015](https://doi.org/10.1017/S0140525X00004015)
- (2006). *Inner presence: Consciousness as a biological phenomenon*. Cambridge, MA: MIT Press.
- (2010). *Consciousness: The science of subjectivity*. Hove, UK: Psychology Press.
- Revonsuo, A., Tuominen, J. & Valli, K. (2015). The avatars in the machine: Dreaming as a simulation of social reality. *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Valli, K. (2011). Dreaming in the multilevel framework. *Consciousness and Cognition*, 20 (4), 1084-1090. [10.1016/j.concog.2011.04.004](https://doi.org/10.1016/j.concog.2011.04.004)
- Valli, K. & Revonsuo, A. (2009). The threat simulation theory in light of recent empirical evidence: A review. *American Journal of Psychology*, 122 (1), 17-38.
- Walker, M. P. & van der Helm, E. (2009). Overnight therapy? The role of sleep in emotional brain processing. *Psychological Bulletin*, 135 (5), 731-748. [10.1037/a0016570](https://doi.org/10.1037/a0016570)
- Windt, J. M. (2010). The immersive spatiotemporal hallucination model of dreaming. *Phenomenology and Cognitive Science*, 9 (2), 295-316. [10.1007/s11097-010-9163-1](https://doi.org/10.1007/s11097-010-9163-1)

Davidson on Believers

Can Non-Linguistic Creatures Have Propositional Attitudes?

Adina Roskies

Donald Davidson has argued that only language-users can have propositional attitudes. His strongest argument in support of this claim is one that links having propositional attitudes to language via a concept of belief. Here I consider various possible interpretations of this argument, looking first at the canonical conception of a concept of belief from the Theory of Mind literature, then at a weaker notion of the concept of belief corresponding to a conception of objective reality, and finally at an intermediate notion involving the ability to attribute mental states. I argue that under each of these various interpretations, analysis and appeal to empirical evidence from developmental and comparative psychology shows the Davidsonian argument to be unsound. Only on a reading of the argument that slides between different interpretations of “concept of belief” are all the premises true, but in that case the argument is invalid. I conclude that Davidson doesn’t provide sufficient reason to deny that non-linguistic creatures can have propositional attitudes.

Keywords

Belief | Capacity | Concept | False belief test | Language | Non-linguistic | Propositional attitudes | Rationality | Thought | Truth

Author

Adina Roskies

adina.l.roskies@dartmouth.edu
Dartmouth College
Hanover, NH, U.S.A.

Commentator

Ulrike Pompe-Alama

ulrike.pompe-alama@philo.uni-stuttgart.de
Universität Stuttgart
Stuttgart, Germany

Editors

Thomas Metzinger

metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Introduction

More often than not, great divides have been postulated between humans and other animals: it has variously been maintained that only humans have souls; that only humans laugh; that only humans play; that only humans are rational. The status of these claims is not merely of theoretical interest: human exceptionalism has long been used to justify or discount arbitrary and often inhumane treatment of animals, including the abuses perpetrated in factory farms and the devastation of habitats for human gain. While the issue of the soul is beyond empirical confirmation or disconfirmation, many

other claims about the uniqueness of humans have been shown to be untrue or only half-true. Recently, in response to philosophical and empirical work, there has been significant political pushback. For example, the Great Ape Project (<http://www.projetoap.org.br/en/>) aims to establish great apes as persons with recognized legal rights. Whether we should stand behind such a project or other less ambitious efforts to treat animals as entities with moral worth depends at least in part on what kind of capacities they have, both cognitive and affective.

Here I combat a philosophically prominent claim of human uniqueness: Donald Davidson's famous argument that only humans can think. In the light of the complex cognitive activities of which animals are clearly capable, one might think this patently untrue. However, Davidson means by this not that animals have no cognitive capacities at all, but that nonhuman animals cannot have beliefs, desires, and other propositional attitudes. What the thesis does is set animal cognition apart from human cognition as a different natural kind, due to radically different representation schemes (see also arguments in [Malcolm 1972](#)). This is not a straw man, but an interesting and challenging thesis. In critically evaluating the arguments Davidson provides in light of empirical evidence from developmental psychology and ethology, insight can be gained into the nature of the relationship between thought and language. Despite its *prima facie* plausibility, I conclude that in light of contemporary studies from human and animal cognition, arguments for restricting propositional attitudes to humans fail.¹ The implications of this result could be far-reaching. Language as a cognitive ability has held a special status in analytic philosophy, where it is often assumed to be foundational to thought and cognition. Rethinking the role of language as a cognitive newcomer and possibly in large part a cognitive overlay resting atop a toolbox of already-powerful cognitive abilities may lead us to rethink a number of fundamental issues in philosophy, as well as to reconsider our cognitive and ethical relationship to the rest of the natural world. This critique of Davidson is illustrative of Dennett's caution:

[p]hilosophy of psychology driven by the concerns of philosophy of language does not fall happily into place. ([Dennett 1987b](#), p. 204)

The various arguments Davidson supplies for thinking that humans are unique in having propositional attitudes all rest upon the idea that

having language is an enabling condition for having propositional attitudes. Since only humans have language, it follows that only humans have propositional attitudes.² Thus, his main argument against animal thought is:

P1 If something has propositional attitudes, then it has language.

P2 Animals don't have language.

C Animals don't have propositional attitudes.

The logic here is unassailable: if [P1](#) and [P2](#) can be established then the conclusion that animals lack propositional attitudes follows. For [Davidson](#), having language is having the ability to speak ([1984](#), p. 167, [2001a](#), p. 99), to express one's thoughts, and to understand the speech and propositional attitudes of others. It is generally accepted that nonhuman animals don't have this ability, despite some evidence that certain birds and higher mammals have some nontrivial linguistic abilities ([Kaminski et al. 2004](#); [Pepperberg 2000](#); [Savage-Rumbaugh 1986](#)). Therefore, we will grant [P2](#).³ The success of this argument denying propositional attitudes to nonhumans therefore rests on the ability to establish [P1](#), namely the claim that having propositional attitudes requires language. In this paper I consider the various avenues by which Davidson tries to establish [P1](#), for his arguments make contact with a broad range of research concerning mind and language, and serve as a good guide to attempts to link propositional attitudes to language. I begin by situating Davidson's arguments in his larger theoretical context, and raise a few methodological worries about his approach. I then briefly consider some of his minor arguments, before turning to his strongest argument linking propositional attitudes to language. I argue that on the most plausible consistent readings of

¹ However, some very interesting and very recent (currently unpublished) work by Susan Carey calls into question the interpretation of some of the extant pro-propositional attitude empirical work.

² Davidson equates thought with propositional attitudes. He famously expresses his denial of propositional attitudes to nonhuman animals as the claim that animals can't think. See [Davidson \(2001a\)](#). Here I focus upon arguments found in his 1975 paper "Thought and Talk", and his later paper "Rational Animals".

³ Another reason for focusing on [P1](#) rather than is that finding counterexamples to [P2](#) will at most make room to usher specific species into the thought-capable fold, but will not challenge Davidson's argument directly.

his arguments, one or another premise can be empirically falsified. I conclude by considering how to proceed to better understand the nature and limitations of animal thought.

2 Initial considerations

2.1 Propositional attitudes

A creature is said to have a propositional attitude when she stands in some appropriate relation (i.e., hoping, wanting, fearing, believing, etc.) to a proposition.⁴ What propositional attitudes are, and who may enjoy them, may well be influenced by what one takes propositions to be. For instance, a skeptic about propositions may deny that anyone has propositional attitudes in the above sense. For our purposes it is not necessary to resolve questions about the nature of propositions, provided that we accept that humans can (and do) have propositional attitudes—meaning that there is something *proposition-like* to which a thinker can be appropriately related, whether this be a sentence (Fodor 1978), a set of possible worlds (Lewis 1979; Stalnaker 1984), or a state of affairs (Marcus 1990). What remains to be determined is whether appropriate relations to proposition-like entities can be supported in non-linguistic creatures.

2.2 Methodological attitudes

Davidson's arguments are offered in the context of his larger theoretical commitments to the nature of mind and meaning, commit-

ments that stem from his interpretationist philosophy.⁵ In general, interpretationist strategies answer the following three questions simultaneously: “In virtue of what does a creature have propositional attitudes?”, “which propositional attitudes do they have?” and “when is one justified in attributing these attitudes to a creature?” According to interpretationism, a creature has propositional attitudes in virtue of being interpretable; the most coherent, charitable interpretation that accurately (or accurately enough) predicts behavior is the justified interpretation; and the contents of that interpretation serve to determine the contents of the creature's propositional attitudes. As Byrne puts it, in an interpretationist strategy, “there is no gap between our *best judgments* of a subject's beliefs and desires and the *truth* about the subject's beliefs and desires,” (1998). Thus, if a creature's behavior can be accurately predicted or explained by an attribution of beliefs and desires in conjunction with the assumption of rationality, we are justified in attributing propositional attitudes to the creature.

Davidson's strongest arguments for why thought requires language are motivated by his interpretationism. On a strict interpretationist view, meaning does not exist without interpretation; so if a system is uninterpreted, it lacks contentful states. Davidson believes that language is a prerequisite for entering the world of interpretation. If no language, then no interpretation, so no content. But let us consider, from an interpretationist stance, why one might think that language is a prerequisite for interpretation.

One might think that Davidson is moved by the idea that only linguistic behavior can be interpreted. However, this cannot be Davidson's position. If it were, Davidson's approach to propositional attitude attribution would be at odds with his own interpretive strategy for attributing content to mental states. The basic idea of Davidson's interpretationism is that in ascribing content to another person's mental states, we

⁴ Some have argued that there is no account of what a proposition is that is both coherent and satisfies the various criteria that propositions are traditionally supposed to satisfy (that tradition stemming initially from Frege). See e.g., Dennett (1987a), and Churchland (1981). It is unfortunate, but true, that if our notion of a proposition is fundamentally incoherent, and no compromises can be reached on the criteria propositions must satisfy, then there is no such thing as a proposition. *A fortiori*, we can't stand in any meaningful relation to propositions, so we lack propositional attitudes. Such is the position of some eliminativists. Others have compromised on the demands put on propositions. Quine, for instance, while being no friend of abstract entities such as propositions as usually conceived, found sentences to be less ontologically troublesome stand-ins for them, and held that to have a propositional attitude is to stand in some relevant relation to an eternal sentence—thereby still satisfying our philosophical intuitions about the role of propositional attitudes in explanations of human thought and behavior.

⁵ In the literature, “interpretationism” is often used interchangeably with “interpretivism”. Since “interpretivism” is more commonly used to denote a strategy of legal interpretation, I will use “interpretationism” here.

assume that that person is rational, and we ascribe content to her utterances, behaviors, and mental states in such a way as to maximize the coherence of that person's beliefs and desires in light of her behavior. Undeniably, there is a class of behaviors that humans have and animals lack, namely linguistic behaviors. However, both humans and animals share a wide range of non-linguistic behaviors that admit of interpretation. On the face of it, those behaviors provide ample evidence upon which to base attributions of mental content, and Davidson himself would not refuse to attribute propositional attitudes to a silent person. However, Davidson pointedly refuses to apply a straightforward interpretationist strategy to non-linguistic animals. To avoid arbitrariness, an independent argument is needed to privilege language over other behaviors.

Perhaps Davidson believes that rationality is impossible without language. If we cannot attribute rationality to a creature, the interpretationist strategy does not apply. More than a few people have argued that animals are not rational, yet there is reason to believe, under some plausible construals of rationality, that they are. To hold that rationality presupposes language commits one to a narrow view of rationality, already colored by a linguistic bias. Such a view implicitly begs the question in which we are interested. Admittedly, what rationality is is a vexed question in philosophy, and determining whether a creature is rational falls prey to the same holistic problems as determining whether it has propositional attitudes. A theory of rationality predicated upon a conception of practical reason instead of upon linguistic manipulation appears to be more neutral. There is abundant evidence for practically rational behavior in the animal world. After all, animals of all stripes are here now because they have been evolutionary successful, and to have succeeded requires in some nontrivial sense that goals are achieved by instrumental behavior.⁶ All animals exhibit some degree of ration-

ality, construed in this way. Building on this view of rationality promises to enable us to posit criteria or hallmarks for minimally rational behavior that are independent of language, yet also to concede that some rational behaviors are linguistically dependent, and thus unique to humans. Indeed, one might think that a good way to assess rationality would be to see to what extent an animal's behavior is predictable or explicable with reference to survival requirements and common sense belief-desire psychology. A wide range of animal behaviors certainly seem apt for explanation with reference to the rational interplay of ecologically-relevant propositional attitudes. If one thinks that aptness for explanation in terms of rationality is sufficient evidence for rationality, and accepts, as Davidson does, that rationality rests on the interplay of propositional attitudes, then we have ample evidence that animals have propositional attitudes, rather than that they do not.

Davidson, however, obviously thinks that the reasons to deny animals propositional attitudes supersede reasons to attribute rationality to them; he applies *modus tollens* to my *modus ponens*. Since he denies that animals have propositional attitudes, and he thinks rationality requires propositional attitudes, he denies that animals are rational. We are led to very different conclusions about the nature of animals' mental lives depending upon whether we take ourselves to be more justified in attributing rational behavior to them or in refusing to attribute to them propositional attitudes. Because the questions of propositional attitudes and of rationality are both equally troubling and closely linked, arguments against animal thought based on assumptions about rationality are not compelling.

Thus, we have as yet failed to find ample reason to refuse to apply the basic interpretationist strategy to non-linguistic animal behavior. Perhaps Davidson thinks that, in the absence of language, we have insufficient evidence for attributing propositional attitudes to animals. Perhaps it is because Davidson thinks that "having the gift of tongues" is both necessary and sufficient for having propositional attitudes (1984, p. 156, 2001a, p. 104), he views language

⁶ Decision theory, for instance, gives us one model of rationality. Interestingly, in many ecological studies of foraging behavior that use decision theory to assess animal choice, animal behavior is found not to just be adaptive, but optimal. For example, animal foraging decisions approach optimality. See e.g., Stephens & Krebs (1986).

possession as *the* evidential criterion for propositional attitude attribution. He consequently denies that we can be justified in attributing propositional attitudes to creatures on the basis of non-linguistic behavior. Language gives the radical interpreter the green light: evidence of linguistic behavior licenses application of the radical interpretive strategy.

Even if we grant that language is the best evidence for propositional attitudes, we should be immediately suspicious of the presumption that the only evidence relevant for deciding whether something has propositional attitudes is the presence of a necessary and sufficient condition for having them. In normal empirical inquiry, criteria that are necessary and sufficient are rarely the only ones that qualify as evidence for assessing empirical claims. For instance, a rash may be relevant evidence for determining whether a person has Lyme disease, despite the fact that not all people with rashes have Lyme disease, and not all people with Lyme have rashes. Might there not be evidence highly indicative of whether a creature has propositional attitudes, despite the fact that the evidence is not decisive? Reasonable, predictable behavior is surely a clear source of evidence for the existence of propositional attitudes, despite the fact that it only provides defeasible reasons for thinking they exist.

Furthermore, unlike instrumentalists like Dennett, Davidson seems to favor the idea that beliefs are real; his anomalous monism posits a physical-causal substrate for mental states, albeit one that exempts psychology from being reduced to physical laws.⁷ One might think, nonetheless, that it would be reasonable for a realist to accept the possibility that beliefs involve some internal representational structures, and that there could therefore be other types of reliable evidence besides linguistic evidence for the presence of propositional attitudes. Thus, Davidson's exclusive focus on language is in tension with his realist leanings. Furthermore, if

one is a realist about thought, it is not the evidential question, but rather the question of the grounds of possibility for having propositional attitudes that should be of primary interest. The Davidsonian mix of interpretationism and realism creates an uneasy tension, for while he tends toward realism about belief, he often seems to think the metaphysical and epistemological construals of the question amount to the same thing: a creature has propositional attitudes if we ought to interpret him as having them. I suspect that this collapsing of the issues accounts for Davidson's view that the question of whether a creature has propositional attitudes is closely tied to the evidential question of what evidence is relevant for deciding whether something has propositional attitudes.

There is, as far as I can tell, a lack of a substantive argument for requiring that a creature has language to be a candidate for interpretation, as well as for holding that only the presence of language provides sufficient evidence for attributing propositional attitudes. Thus, neither the interpretationist strategy itself, nor Davidson's concerns about evidential warrant justify the position that only language-speaking creatures can be candidates for propositional attitudes. Now let us turn to the specific arguments Davidson offers for denying animals propositional attitudes: the reasons he offers for holding that language is necessary for thought.

3 Minor arguments

Why might someone think that language is necessary for having propositional attitudes? A common reason for supposing that language is necessary for thought is that one is in the grip of a picture about the nature of thought—namely that thought is a type of language, or is linguistic or language-like. If propositions are linguistic entities, then creatures that lack the capacity for linguistic representation might well be unable to represent propositions and thus be unable to hold an attitude toward a proposition. However, since there are competing accounts of what propositions actually are, several of which see them as non-linguistic in nature, the intuitive language-like characteristics of pro-

⁷ The debate about propositional attitudes, language, and capacity for thought has implications beyond philosophy of mind to ethics. As Davidson himself noted, personal and sub personal levels of description refer to different logical subjects, and thus Davidson's argument has implications for the possibility of attributing *personhood* to animals. See again, <http://www.projeto-gap.org.br/en/>.

positions does not settle the question (Lewis 1979; Stalnaker 1984).

In several places Davidson gestures at related arguments for denying non-linguistic creatures propositional attitudes (1984, p. 156, 2001a, p. 98). These stem from an implicit commitment to propositional attitudes having certain characteristics that only languages possess. For instance, Davidson claims that propositional attitudes have definite content, and that only things expressed in language have definite content. Drawing on the discussion of Malcolm (1972) before him, he gives an example of a dog chasing a cat up a tree. Like Malcolm, he notes that we cannot attribute to the dog the thought that the cat ran up the maple, as opposed to that the cat ran up the tree. If there is no particular thought we can attribute to the dog, then the dog hasn't had a thought with definite content, and so hasn't had a propositional attitude. Davidson elsewhere claims that propositional attitudes are opaque⁸, and that language accounts for their opacity (Davidson 2001a, p. 97). Although these claims can be combatted directly, I will not pursue those arguments here. Both the definite content claim and the opacity claim lose their teeth when it is recognized that they take the following form:

P1	Propositional attitudes have a property, p
P2	Language has property p
<hr/>	
C	Therefore, language is necessary for propositional attitudes

This argument is fallacious—it would only be valid if nothing *but* language had property p. But no such argument is on offer. It is worth noting, moreover, that whether propositional attitudes have the property p in question is itself contentious—do all our beliefs have definite content? Finally, even if having property p were somehow constitutive of thought, and to have p thought had to be linguistic, this would still not entail that a creature with beliefs and desires must have language in Davidson's sense. Fodor

(see Fodor 1975), for instance, thinks that a creature must have a language of thought to have propositional attitudes, but he holds that it need not be able to speak or understand a public language to have a language of thought. Even if claims about definite content and opacity were true, that is, if Fodor is right, Davidson has erred in thinking that thought requires an external as opposed to an internal language. If animals have a language of thought, they are non-language-using believers.

4 Davidson's Master Argument

The above minor arguments don't play a central role in Davidson's support of P1. The strongest support for the crucial premise is found in what I will call his Master Argument.⁹ The Master Argument puts psychological restrictions on what it is to be an interpreter, and it supports the claim that one cannot have propositional attitudes without language. If the Master Argument succeeds, then Davidson's arguments for denying that animals have propositional attitudes is compelling. But, as I shall argue, the Master Argument ultimately fails, and thus also fails to support the denial of propositional attitudes to animals.

According to Davidson's interpretationism, having beliefs entails being an interpreter. The basic idea of the Master Argument is that possessing certain concepts is a prerequisite for being an interpreter, and that an organism must have language in order to have these concepts.¹⁰

⁹ Davidson nowhere presents his Master Argument in this precise form. I reconstruct the logical form of his argument from "Thought and Talk" and "Rational Animals".

¹⁰ This ought to be distinguished from the idea that having propositional thought requires having some concepts, and that the contents that can be entertained by a creature in propositional thought are constrained by the set of concepts that the creature possesses. This view, held by a variety of thinkers from Frege to Fodor, stems from the belief that the propositions to which a thinker stands in relation in having a propositional attitude are complex entities composed of concepts. But then the question of whether animals have propositional thought can be recast as the question of whether animals have concepts. If, additionally, one combined this view of the cognitive structure of propositions with a view according to which concept possession requires language, one would have an argument for why language is necessary for propositional thought. However, whether concept possession requires language is a question that depends, among other things, on what concepts are. Whether the vehicles of thought are language-like, as I argued earlier, is orthogonal to the issue of whether an organism possesses the capacity to speak or understand speech. Therefore Davidson's argument cannot rest on the nature of concepts.

⁸ Substitution of co-referring terms in "opaque" contexts may not preserve truth. Such is the case with propositional attitudes. Thus, while it is true that Lois Lane believes "Superman is a hero", it may be false that she believes "Clark Kent is a hero", despite the fact that Clark Kent is identical to Superman.

Davidson's position differs from the more widely-held view that having *some* concepts is required for having propositional thought, by supposing that there are *specific* concepts that a creature must possess in order to have propositional thought. The Master Argument links thought to language by way of higher-order thoughts. Specifically, Davidson suggests that a concept of *belief* is a prerequisite for propositional attitudes, and that a concept of belief is unavailable without language. Here is Davidson's Master Argument:

M1 If S has propositional attitudes, then S has beliefs.

M2 If S has beliefs, then S has a concept of belief.

M3 If S has a concept of belief, then S has language.

MC If S has propositional attitudes, then S has language.

The argument is clearly valid. But is it sound?

M1 is plausible; it just highlights Davidson's view that beliefs are a fundamental propositional attitude, and that to have any propositional attitudes at all, a creature must have some beliefs. **M2** and **M3**, the remaining premises, are interesting, but their meaning is unclear, for they contain a clause that needs to be unpacked: what exactly is a "concept of belief"? Let us distinguish three different conceptions of a "concept of belief", varying in stringency. One conception of the concept of belief is robust, in which the concept of belief is the fully articulated belief-concept that is taken to be definitive of a mature theory of mind. On this robust view, having a concept of belief is an epistemologically-rich notion that entails having an ability to pass the "false belief test". That is, it is criterial for having the concept of belief that one has the ability to attribute to others a mental representation of the world that may differ from the way the world is, as well as a recognition of the perceptual circumstances that would engender false representations. In contrast, a deflationary conception of what it is to have the concept of belief merely requires an understanding that the world is distinct from how it appears

or how one takes it to be, or that belief can come apart from reality. On a deflationary view, then, having the concept of belief is rather like having the concept of an objective reality. Finally, we might consider an intermediate notion of the concept of belief that involves the ability to attribute representational mental states to oneself and others, without satisfying all the constraints that a robust conception must meet. Which, if any, of these conceptions of "concept of belief" is important for Davidson's argument linking belief to language?

4.1 The robust conception of belief

The robust conception of belief became important in developmental psychology in the context of concerns about Theory of Mind: having a notion of false belief was taken to be diagnostic of a mature TOM, and, according to many researchers in the field, only develops in humans at around four years of age (Saxe et al. 2004; Wellman et al. 2001; Wimmer & Perner 1983). However, requiring a concept of belief in the robust sense seems too demanding a condition for having propositional attitudes. While we might plausibly doubt whether prelinguistic infants really have propositional attitudes, it is hard to deny that young children who have already acquired a sophisticated facility with language have propositional attitudes. Children of two and three, for instance, clearly refer to objects in the world using language, and they readily express their desires ("I want the green monkey!"), beliefs ("I think the ball is under the bed"), as well as fears and other propositional attitudes. They understand others, refer to their own and others' mental states, and communicate effectively. We typically and with great conviction attribute propositional attitudes to children of these ages. Nonetheless, according to most developmental psychologists (See e.g., Perner et al. 1987; Call et al. 1999; http://youtu.be/8hLubgpY2_w), until the age of four (two years after they develop considerable language abilities) children lack a concept of belief in the robust sense.¹¹ And if so, we

¹¹ Kristen Andrews takes autistic subjects to be counterexamples to Davidson's view, which would also argue against **M2**. (Andrews 2002).

make ordinary propositional attributions to children well before they possess the robust concept of belief. Thus, [M2](#) is false.¹²

Not all psychologists agree that a robust concept of belief doesn't develop until about four years of age. Some have argued that the methods used in many of the classic false belief studies rely too much on language or on inhibitory control, and that tests other than the classic false belief test are sufficient for demonstrating understanding of false beliefs. For instance, a recent study suggests that children have a concept of belief at far earlier ages than previously thought—earlier, in fact, than the development of language ([Onishi & Baillargeon 2005](#); [Baillargeon et al. 2010](#); [Caron 2009](#)). However, if this is so, then [M3](#) is false, for the robust conception of belief does not depend on having language. This version of the Master Argument depends upon a tight connection between competence in the false-belief task and belief. On one conception of what evidence is sufficient to reflect performance on the false-belief task, [M2](#) is false, and on another conception, [M3](#) is false. Either way, the Master Argument is empirically refuted, and the robust conception of “the concept of belief” fails to link language possession and propositional attitudes.¹³

Davidson may well be unperturbed, for there is no textual evidence that he means to implicate the robust conception of belief when he claims the concept of belief is necessary for having beliefs. After all, from the standpoint of his radical interpreter, one can only be a believer in virtue of interpreting others, but it is unclear why the possibility of such interpretation should rest upon a grasp of others' mental states being beliefs in this robust sense, rather than in some weaker sense. In “[The Second Person](#)” (1992), for instance, [Davidson](#) argues that for our mental

states to have determinate content we must interact with another being in order to “triangulate” and thus make determinate the referents of our thoughts. Nothing in this picture requires that an interpreter have a robust concept of belief as opposed to a more deflationary one.

4.2 The deflationary conception of belief

In line with the idea that Davidson has a more deflationary view in mind, in both “[Thought and Talk](#)” and “[Rational Animals](#)” he mentions a different criterion for having a belief, which he also thinks links the possession of language to the ability to have propositional attitudes. This is the criterion of possessing a concept of “objective truth.” Davidson's argument for language via the criterion of objective truth is as follows:

O1 In order to have propositional attitudes, one must have beliefs.

O2 In order to have beliefs, one must have a concept of objective truth.

O3 In order to have a concept of objective truth, one must have language.

OC Propositional attitudes require language.

The logic here is again unproblematic, but unpacking the premises is not. At times Davidson seems to equate the concept of objective truth with that of belief. I take this as evidence that he intends “the concept of belief” in the Master Argument in its most deflationary interpretation: as an understanding that how the world is can come apart from how one takes the world to be. Given this interpretation one could believe that the concept of objective truth co-occurs with that of belief, or that the cognitive conditions that make possible the concept of belief are the same as those that make possible the concept of objective truth. In any case, Davidson sees a tight connection between the notions of belief and objectivity.

How are we to understand the “concept of objective truth” in [O2](#) and [O3](#)? If Davidson means it to be a metasemantic concept, such as having a Tarskian definition of truth, or an understanding that truth applies to propositions, and so on, then it would be almost assured that

¹² In addition, at ages far younger than those at which children pass the false-belief task, they act as interpreters, in Davidson's sense. Any parent knows that their children interpret speech well before they are speakers, and long before the age at which they pass the false-belief task. So if interpretation is central to having propositional attitudes, it doesn't require a robust theory of mind.

¹³ Of course, Onishi and Baillargeon's interpretation is subject to refutation. Should their findings (they developed a nonverbal task that suggest that infants much younger than previously supposed represent others' mental states, such as goals, perceptions and beliefs.) reflect something like proto-beliefs rather than full-blown propositional attitude-sustaining beliefs, [M3](#) would not be falsified.

one could not grasp the concept of truth without language. It would explain the *prima facie* plausibility of the Objective Truth version of the Master Argument. However, if we adopt that reading of objective truth, O2 would be false, for people certainly have propositional attitudes even if they never become philosophers, and even if they never have an inkling about metasegmental notions.

Another clue about what Davidson means by objective truth comes from his emphasis on triangulation. Davidson thinks we need to interact with another person in order to come to see the world as external to us—in order to develop a notion of objectivity. By linguistically triangulating on an object with another, we are forced to recognize that object as part of an objective reality. Davidson illustrates this view in “The second person”:

Belief, intention, and the other propositional attitudes are all social in that they are states a creature cannot be in without having the concept of intersubjective truth, and this is a concept one cannot have without sharing, and knowing that one shares, a world, and a way of thinking about the world, with someone else. (2001b, p. 121)

However, there are two fundamental problems with using triangulation as an argument for the necessity of language for thought. First, there is nothing apparent about triangulation that requires spoken language as opposed to some other sort of joint interaction or non-linguistic communication. It is, indeed, difficult to see why language as opposed to action would be operative in developing a notion of a world external to ourselves. So triangulation fails to show that language is necessary for thought. Second, it is difficult to see how triangulation could itself suffice for a notion of objectivity. In order for me to triangulate with another, I must *first* see the other as part of the external world, as opposed to an element in my mentality. As long as the other is merely a part of the way I take things to be, it cannot fulfil the role of the second person (see for example, Roskies 2011).

So triangulation also fails as a mechanism for constructing the concept of objectivity. Nonetheless, Davidson’s emphasis on triangulation strongly suggests that by “objective truth” he is referring to the appearance/reality distinction.

This interpretation is further strengthened by taking seriously the fact that Davidson thinks the concepts of belief and truth are closely linked (1984). As mentioned earlier, having the concept of objective truth is nothing other than understanding that how the world is can come apart from how one takes the world to be. What evidence do we have that language is required for this?

4.3 Surprise

As further evidence that Davidson intends a deflationary view of the concepts of belief and objective truth, we can turn to another formulation of the Master Argument. In his most forthright explication of what he means by “concept of belief”, he suggests that there is a behavioral mark that is coextensive with having such a concept: surprise.

In order to have any propositional attitude at all, it is necessary to have the concept of a belief, to have a belief about some belief. But what is required in order to have the concept of a belief? Here I turn for help to the phenomenon of surprise, since I think that surprise requires the concept of belief. (Davidson 2001a, p. 104)

The willingness to consider some sort of non-linguistic behavior as relevant to the question of whether a creature has propositional attitudes is a methodological breakthrough, for it provides an avenue independent of language for assessing whether an animal has the requisite cognitive machinery to be a believer. Davidson maintains that the ability to be surprised is diagnostic of having the concept of belief. It indicates recognition that one’s own mental representation fails to conform to that which it represents, and as such it constitutes necessary and sufficient evidence of the concept of belief.

Following this intuition, we can amend Davidson's Master Argument to incorporate this insight:

S1 If S has propositional attitudes, then S has beliefs.

S2 If S has beliefs, S has a concept of belief.

S3 S has a concept of belief iff S has the capacity for surprise.

S4 If S has the capacity for surprise, S has language.

SC Propositional attitudes require language.

The idea that surprise goes hand-in-hand with the concept of belief is not implausible: if surprise issues from the recognition that one's belief about how the world is fails to correspond with the way the world is, then surprise is good evidence for the concept of belief. Moreover, because this idea does not have implications for the ability to attribute propositional attitudes to others in an operative sense, it suggests that the interpretation of "concept of belief" that Davidson favors is a deflationary interpretation: one that involves appreciation of the appearance/reality distinction, or, as discussed above, the concept of objective truth. Thus, S2 takes the deflationary interpretation of the concept of belief, and for the argument to be valid, S3 must also take that interpretation.

Unfortunately for this version of the argument, S4 is false. There is clear and abundant empirical evidence that the ability to be surprised at the mismatch between the world and one's own representation of the world is independent of language (Dupoux 2001; Feigenson et al. 2002; Hauser & Carey 1998; Santos et al. 2002; Wynn 1992). Take, for example, an invaluable tool in the developmental psychologist's toolkit: the violation of expectancy looking method (V) for testing infants. Many studies performed on pre-linguistic human infants employ this paradigm in order to explore what an infant knows. The idea is simple: infants look longer at stimuli that fail to correspond with their expectations. This method has been used to determine, among other things, that infants have an innate (or very early developing) concept of number. In now classic experiments,

Wynn and colleagues demonstrated that infants can do simple arithmetic (Wynn 1992). She showed infants as young as five months a toy, and placed it behind a screen. Then she showed them another toy and also placed it behind the screen. The screen was then lowered, revealing either two toys (the expected outcome), or only one toy. Infants looked longer at the unexpected outcome. The same paradigm was used with different numerical combinations, demonstrating that for numerosity up to three, infants can do simple addition and subtraction, and are surprised when what is revealed behind the screen does not comply with their expectations. Significantly, this robust effect, which is due to surprise, precedes the development of language by more than a year.

Davidson might reply that it is not actually possessing language, but rather possessing the *capacity for language* that is important for surprise, and thus for the concept of belief. Maybe, even though they cannot yet speak, infants possess a language faculty, which, immature as it may be, is sufficient to support surprise. However, this attempt to patch the argument also fails. The VELM is used frequently in studies with nonhuman primates, and while they never develop language nor seem to have a capacity for natural language, they too exhibit surprise when their expectations are violated (Hauser 2000; Hauser et al. 1996). So, it seems, language is not a requirement for surprise, nor is surprise evidence for the presence of or capacity for language.

The empirical studies of developmental psychologists and primatologists undermine the Surprise version of the Master Argument: surprise does not depend upon having language. Moreover, if premises S2 and S3 are true—if the capacity for surprise is evidence of the concept of belief, and if propositional attitudes depend upon possession of the concept of belief—then propositional attitudes do not depend upon language.

Let us briefly revisit the Objective Truth version of the Master Argument. I have argued that only a deflationary notion of objective truth is a candidate interpretation for the argument. I have also suggested that this is the only

notion of “the concept of objective truth” that meshes with the arguments Davidson raises regarding belief and surprise. Thus, having a concept of objective truth is having a concept that the way the world is can come apart from how one takes it to be. If this is correct, then the Objective Truth version of the Master Argument is false.

In Wynn’s looking-time studies discussed above, the child has clearly developed expectations of what lies behind the screen, and must somehow represent this to herself. When the screen is lowered and the child sees what is behind the screen, there must be some sense in which correspondence with the expectation or lack of correspondence is noted, and in which the data coming in from the senses is privileged over the internal representation. This is, in essence, what it is to recognize that beliefs about the world can come apart from the way the world is. Clearly this sort of grasp of reality does not depend upon language: pre-linguistic infants and non-linguistic animals possess it. One can easily imagine how violation of expectation can be instantiated in a system with imagistic thought. The languageless child need only conjure up an image of the objects behind the screen and compare this with the visual scene before him. As long as the child privileges the sensory information over the mental representation, we might say that he has a concept of reality and of the belief/reality distinction. In summary, then, language is not required for a concept of objective truth.

4.4 The intermediate conception of belief

We have ruled out both the robust and weakest notions of “concept of belief” as candidate notions for a successful interpretation of Davidson’s argument linking belief to language. Perhaps an intermediate notion can do the job. This notion involves the ability to attribute representational states to oneself and others; it is less sophisticated than that required to pass the false-belief task, but more complex than the recognition of an appearance/reality distinction.

One potential reason why representational-state attribution may be important for having

beliefs involves self-reflection: perhaps being a believer requires being able to think of oneself as a believer, and thus requires the concept of belief. This amounts to the claim that beliefs cannot be held non-reflectively. Since we clearly do have beliefs that we do not have beliefs about, what is at issue is not the actuality of having beliefs about beliefs, but the possibility or capacity to do so. However, while there are arguments that the ability to think about oneself as a believer is required for a rich construal of theoretical rationality (see [Bermúdez 2003](#), Ch. 7), there is no clear argument why such reflective ability should be constitutive of having beliefs. Indeed, it seems like the ability to believe things about one’s beliefs would require that one could believe things, so that belief is conceptually prior to self-reflection. In any case, self-reflection is not Davidson’s stated reason for thinking that the concept of belief is important for having beliefs.

The other reason to hold that having belief requires having a concept of belief under the intermediate conception links the ability to attribute mental states to others with having the concept of belief. Thus there are two different strengths of intermediate interpretations to consider. According to the less demanding interpretation, a concept of belief is required in order to attribute contentful states to other creatures; whereas the more demanding interpretation holds that a concept of belief is required to attribute propositional attitudes to others: one must be an interpreter, not just an interpretee.

We can discount the less demanding of these interpretations for the purpose of this argument linking thought to language,¹⁴ because if [M2](#) (“If S has beliefs, then S has a concept of belief”) is interpreted in this way, then [M3](#), the claim that language is required for a concept of belief, read in this way, is false. There is growing evidence that non-language-using animals are able to attribute representational states to other animals. One compelling illustration of this comes from ([Hare et al. 2000](#)), who show

¹⁴ We ought to reject this interpretation for the purposes of Davidson’s argument, despite the fact that we may ultimately agree with it as a necessary condition for having propositional thought.

that subordinate rhesus monkeys only approach food in the presence of a dominant male when they know that the male is unable to see the food (interestingly, dominant males appear not to care whether or not a subordinate male sees food, pointing to yet a further level of sophistication in the cognitive processes of non-linguistic animals). Thus, if it is the case that to believe requires having the ability to attribute contentful mental states to others, then it is not the case that believing requires language. Indeed, recent work on non-human primate theory of mind suggests that monkeys and chimpanzees have a theory of mind that represents goal states and distinguishes between knowledge and ignorance of other agents (the presence and absence of contentful mental representations), even if it fails to account for misrepresentation (Call & Tomasello 2008; Kaminski et al. 2008; Martcorena et al. 2011). Although they may have a less articulated theory of mind than we do, we may nonetheless adequately characterize their representational system with mental-state terms (Butterfill & Apperly 2013; Martcorena et al. 2011).

What remains is the notion that the ability to attribute beliefs qua propositional attitudes to others is necessary for having beliefs. That is, not only must they attribute mental states to others, but those mental states must possess the characteristics of beliefs. Remember that we have already discounted the robust notion of belief as too demanding, so what is necessary is not that animals have a notion of false belief per se, but rather that they have a notion of a belief as a representational mental state that can play a role in behavioral explanation or prediction. So far there is no compelling evidence that nonhuman animals have this, consistent with the possibility that such a representational ability as this may indeed require language, or at least some sophisticated ability to symbolize abstractions and predicate them of objects. Whether this is so is ultimately an empirical question. However, at least some philosophers think monkeys may be able to do this. As Lurz characterizes the above studies, animals do have the ability to represent propositional mental states in others—not as attitudes

aimed at representing objective truth, but instead as attitudes with propositional contents that provide information regarding motivation to act (2011a). Lurz characterizes this as a kind of belief–desire attribution. Baillargeon’s data proves relevant here too, for her results are best explained by taking the infants in her study as postulating representational mental states of the actor in order to predict her behavior; violation of their expectation causes them to look longer. Thus, without imputing these infants some understanding of others’ representational mental states, we would be unable to account for this data. However, in this case M3 would then be false, for the linguistic abilities of fifteen-month-old infants typically are minimal—certainly not of the sophistication we would expect would be necessary to linguistically encode a belief-concept. While the evidence that bears on this case is perhaps the least well-established, and this study involves infants at an age when they are poised to develop language, the burden of proof is shifted to the person who wants to argue that language is necessary for a concept of belief. That burden is not discharged: Davidson lacks a positive argument for why this relatively demanding notion of attributing content to others is the one required for an organism to be a believer.

5 Beyond interpretationism

Davidson argues that language is required for thought. His Master Argument posits that having a concept of belief is a necessary intermediary for having propositional attitudes, and that language is necessary for having a concept of belief. Of the various conceptions of “concept of belief” that might play a role in Davidson’s argument, the robust conception is too strong, and empirically falsified. While the robust conception may require language, we attribute propositional attitudes before children are clearly in possession of such concepts. Davidson’s examples and arguments support only deflationary interpretations of the concept of belief and the associated concept of objective truth: those that involve distinguishing between appearance and reality, or those that involve attributing

mental content. However, as numerous studies in developmental and comparative psychology have shown, the deflationary conception is one that many creatures without language enjoy. Even an intermediate conception does not seem to play the role Davidson's argument requires, for the ability to attribute mental content does not require language, and neither does the ability to attribute to others representational mental states, though here the evidence is less clear. Davidson's arguments seem compelling because their plausibility relies upon a slide between less and more demanding conceptions of the concept of belief. For instance, a weak conception of "concept of belief" in [M2](#) and a robust one in [M3](#) yields an argument with apparently true premises, but because the argument equivocates on "concept of belief", the argument is invalid. This analysis, as well as an appreciation of the methodological considerations for using non-linguistic behavior as evidence of propositional attitudes, supports the view that some mental states of non-linguistic animals can aptly be classified as propositional attitudes.

In empirical circles it seems to be taken for granted that at least some non-linguistic animals have mental states best described as propositional attitudes. But this acceptance is merely the first step in a larger project. For example, even if there is good reason to think that non-linguistic creatures have propositional attitudes, how they could have these remains to be elucidated. That is, what is the nature of the representational resources available to them? And given these representational resources, what sorts of contents are they capable of representing? What kinds of reasoning and inference could such representations support? What are the cognitive limitations necessitated by their representational architectures? One can begin addressing these fascinating questions empirically either at the functional psychological level or at the level of representation, and from either level one can work toward answering questions about the other.

Instead of thinking that language itself is what makes complex, structured, or propositional thought possible, we should consider: 1) how non-linguistic capacities could underlie

complex representational abilities 2) the unique elements of linguistic competence and what they may or may not make possible vis-à-vis thought. In an example of the first, [Proust \(1999\)](#); see also [this collection](#)) provides an illuminating philosophical discussion of structured non-linguistic representational abilities (or "structured competences") and how they could make possible objective representations. Structured representations as such could form the building blocks of propositional attitudes. Bermúdez argues for abilities and for certain logical limitations on both the inferential and representational abilities of non-linguistic representers ([Bermúdez 2003](#)). Whether such limitations necessarily obtain is a matter of dispute ([Lurz 2007](#)).

When considering how linguistic abilities could augment thought, it is useful to identify elements of language that could contribute to representational complexity even if present without all the components of language. For example, [Clark](#) suggests that the human language-like ability to use symbols to represent abstract objects allows us to objectify our own thoughts and operate upon them ([2000](#)). Depending on what things can be symbolized, this could make possible metacognition or higher-order thought that might not otherwise be possible. Thus, the ability to represent symbolically can influence the kinds and complexity of reasoning available to a creature, even if that creature is not linguistic in Davidson's sense. Symbolic capacities are necessary but not sufficient for linguistic competence, and could be present even when language is not. And if mere use of symbols is taken to be sufficient for language, then some nonhuman primates are capable of language and thus again can have propositional attitudes. Indeed, it is clear that some nonhuman primates can be trained to use abstract symbols, even if they do not do so naturally. Boysen and colleagues, for example, relate how naïve chimps fail to learn to make second-order generalizations about object classification, but those trained to associate objects with symbols (for relations of "same" and "different") are able to succeed on a second-order classification task ([Thompson et al. 1997](#)). These interesting res-

ults give causal punch to the notion that symbolic objectification is a prerequisite for higher-level or abstract thought, and help to explain the competences that appear to come along with linguistic abilities.

Focusing less on the vehicles and more on the ways in which they can be exploited, Fitch and colleagues argue that recursion, which is a core element of natural language processing, can only operate on symbolic structures subject to rules, and that neither rules nor the objects on which they operate can exist without language-like representations (Hauser et al. 2002). If so, one might expect that forms of reasoning that rely on recursion may only be possible for creatures that also possess linguistic capacities. Thus, use of symbols and recursive rules are two candidates that could help explain the different representational capacities of linguistic and non-linguistic creatures.

6 Conclusion

Here I have argued that Davidson's arguments that nonlinguistic creatures lack thought are either unsound or invalid. While this negative project does not allow us to conclude that they have propositional attitudes or thoughts, it makes room for positive arguments that will take advantage of recent and future empirical work on animal cognition and on the nature of nonlinguistic representations and their role in cognitive processing, as well as for novel negative arguments that might set limits on the capacities of nonlinguistic creatures. Much current research in animal cognition focuses on whether animals have theory of mind paralleling that of humans (Martcorena et al. 2011), or metacognition (Bermúdez 2003; Carruthers 2008; Lurz 2007, 2011a, 2011b; Proust 2010). One might therefore think that the debate has not progressed much since Davidson asked the question about whether animals can have a concept of belief. But Davidson's interest in these questions was narrow, driven by his interpretationism and the view that these states are necessary for being an interpreter and thus for possessing mental content. In contrast, contemporary research does not aim to disprove the existence of

propositional attitudes, but rather to elucidate the scope of these attitudes and understanding the ways in which they may be limited by limitations in representational resources. In the most exciting work, the philosophical and psychological projects come together. This interdisciplinary approach takes seriously evolutionary relationships and has a more nuanced view of the human being's place among other animals. The arguments that result will be of great interest to philosophers of language and mind, as well as to those interested in ethical issues that transcend academia. And while they may vindicate a certain kind of human exceptionalism, they may also articulate our place on a spectrum that will ultimately lead to a more integrated and humane picture of our place in the world.

References

- Andrews, K. (2002). Interpreting autism: A critique of Davidson on thought and language. *Philosophical Psychology*, 15 (3), 317-332. [10.1080/09515080210000061111](https://doi.org/10.1080/09515080210000061111)
- Baillargeon, R., Scott, R. M. & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences*, 14 (3), 110-118. [10.1016/j.tics.2009.12.006](https://doi.org/10.1016/j.tics.2009.12.006)
- Bermúdez, J. (2003). *Thought without words*. Oxford: Oxford University Press.
- Butterfill, S. A. & Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind & Language*, 28 (5), 606-637. [10.1111/mila.12036](https://doi.org/10.1111/mila.12036)
- Byrne, A. (1998). Interpretivism. *European Review of Philosophy*, 3, 199-223.
- Call, J. & Tomasello, M. (1999). A nonverbal false belief task: The performance of children and great apes. *Child Development*, 70 (2), 381-395. [10.1111/1467-8624.00028](https://doi.org/10.1111/1467-8624.00028)
- (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, 12 (5), 187-192. [10.1016/j.tics.2008.02.010](https://doi.org/10.1016/j.tics.2008.02.010)
- Caron, A. J. (2009). Comprehension of the representational mind in infancy. *Developmental Review*, 29 (2), 69-95. [10.1016/j.dr.2009.04.002](https://doi.org/10.1016/j.dr.2009.04.002)
- Carruthers, P. (2008). Meta-cognition in animals: A skeptical look. *Mind & Language*, 23 (1), 58-59. [10.1111/j.1468-0017.2007.00329.x](https://doi.org/10.1111/j.1468-0017.2007.00329.x)
- Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *The Journal of Philosophy*, 78 (2), 67-90.
- Clark, A. (2000). *Mindware: An introduction to the philosophy of cognitive science*. Oxford, UK: Oxford University Press.
- Davidson, D. (1984). Thought and talk. *Inquiries into Truth and Interpretation* (pp. 155-170). Oxford, UK: Clarendon Press.
- (2001a). Rational animals. *Subjective, intersubjective, objective* (pp. 95-106). Oxford, UK: Clarendon Press.
- (2001b). The second person. *Subjective, intersubjective, objective* (pp. 107-121). Oxford, UK: Clarendon Press.
- Dennett, D. C. (1987a). Beyond belief. *The intentional stance* (pp. 117-202). Cambridge, MA: MIT Press.
- (1987b). *The intentional stance*. Cambridge, MA: MIT Press.
- Dupoux, E. (Ed.) (2001). *Language, brain, and cognitive development: Essays in honor of Jacques Mehler*. Cambridge, MA: MIT Press.
- Feigenson, L., Carey, S. & Spelke, E. S. (2002). Infants' discrimination of number vs. continuous extent. *Cognitive Psychology*, 44 (1), 33-66. [10.1006/cogp.2001.0760](https://doi.org/10.1006/cogp.2001.0760)
- Fodor, J. A. (1975). *The language of thought*. New York, NY: Crowell.
- (1978). Propositional attitudes. *Monist*, 61 (4), 501-523. [10.5840/monist197861444](https://doi.org/10.5840/monist197861444)
- Hare, B., Call, J., Agnetta, B. & Tomasello, M. (2000). Chimpanzees know what conspecifics do and do not see. *Animal Behavior*, 59 (4), 771-785. [10.1006/anbe.1999.1377](https://doi.org/10.1006/anbe.1999.1377)
- Hauser, M. D. (2000). *Wild minds*. New York: Henry Holt and Co.
- Hauser, M. D. & Carey, S. (1998). Building a cognitive creature from a set of primitives. In D. D. Cummin & C. Allen (Eds.) *The evolution of mind* (pp. 51-83). Oxford, UK: Oxford University Press.
- Hauser, M. D., MacNeilage, P. & Ware, M. (1996). Numerical representations in primates. *Proceedings of the National Academy of Sciences of the United States of America*, 93 (4), 1514-1517. [10.1073/pnas.93.4.1514](https://doi.org/10.1073/pnas.93.4.1514)
- Hauser, M. D., Chomsky, N. & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298 (5598), 1569-1579. [10.1126/science.298.5598.1569](https://doi.org/10.1126/science.298.5598.1569)
- Kaminski, J., Call, J. & Fischer, J. (2004). Word learning in a domestic dog: Evidence for "fast mapping". *Science*, 304 (5677), 1682-1683. [10.1126/science.1097859](https://doi.org/10.1126/science.1097859)
- Kaminski, J., Call, J. & Tomasello, M. (2008). Chimpanzees know what others know, but not what they believe. *Cognition*, 109 (2), 224-234. [10.1016/j.cognition.2008.08.010](https://doi.org/10.1016/j.cognition.2008.08.010)
- Lewis, D. (1979). Attitudes de dicto, de se. *Philosophical Review*, 88 (4), 513-543. [10.2307/2184843](https://doi.org/10.2307/2184843)
- Lurz, R. W. (2007). In defense of wordless thoughts about thoughts. *Mind and Language*, 22 (3), 270-296. [10.1111/j.1468-0017.2007.00309.x](https://doi.org/10.1111/j.1468-0017.2007.00309.x)
- (2011a). Belief attribution in animals: On how to move forward conceptually and empirically. *Review of Philosophy and Psychology*, 2 (1), 19-59. [10.1007/s13164-010-0042-z](https://doi.org/10.1007/s13164-010-0042-z)
- (2011b). *Mindreading animals: The debate over what animals know about other minds*. Cambridge, MA: MIT Press.
- Malcolm, N. (1972). Thoughtless brutes. *Proceedings and Addresses of the American Philosophical Association*, 46, 5-20. [10.2307/3129585](https://doi.org/10.2307/3129585)
- Marcus, R. B. (1990). Some revisionary proposals about belief and believing. *Philosophy and Phenomenological Research*, 50 (supplement), 133-153.

- Martcorena, D. C. W., Ruiz, A. M., Mukerji, C., Goddu, A. & Santos, L. R. (2011). Monkeys represent others' knowledge but not their beliefs. *Developmental Science*, 14 (5), 1406-1416. [10.1111/j.1467-7687.2011.01085.x](https://doi.org/10.1111/j.1467-7687.2011.01085.x)
- Onishi, K. H. & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, 308 (5719), 255-258. [10.1126/science.1107621](https://doi.org/10.1126/science.1107621)
- Pepperberg, I. M. (2000). *The alex studies*. Cambridge, MA: Harvard University Press.
- Perner, J., Leekam, S. R. & Wimmer, H. (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology*, 5, 125-137. [10.1111/j.2044-835X.1987.tb01048.x](https://doi.org/10.1111/j.2044-835X.1987.tb01048.x)
- Proust, J. (1999). Mind, space and objectivity in non-human animals. *Erkenntnis*, 51 (1). [10.3389/fpsyg.2013.00145](https://doi.org/10.3389/fpsyg.2013.00145)
- (2010). Metacognition. *Philosophy Compass*, 5 (11), 989-998. [10.1111/j.1747-9991.2010.00340.x](https://doi.org/10.1111/j.1747-9991.2010.00340.x)
- (2015). The representational structure of feelings. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Roskies, A. L. (2011). Triangulation and objectivity: Squaring the circle? In C. M. Amoretti & G. Preyer (Eds.) *Triangulation: From an epistemological point of view* (pp. 97-102). Frankfurt am Main, GER: Ontos Verlag.
- Santos, L. R., Hauser, M. D. & Spelke, E. S. (2002). Domain-specific knowledge in human children and nonhuman primates: Artifacts and foods. *The cognitive animal: Empirical and theoretical perspectives on animal cognition* (pp. 205-215). Cambridge, MA: MIT Press.
- Savage-Rumbaugh, S. (1986). *Ape language: From conditioned response to symbol*. New York: Columbia University Press.
- Saxe, R., Carey, S. & Kanwisher, N. (2004). Understanding other minds: Linking developmental psychology and functional neuroimaging. *Annual Review of Psychology*, 55 (1), 87-124. [10.1146/annurev.psych.55.090902.142044](https://doi.org/10.1146/annurev.psych.55.090902.142044)
- Stalnaker, R. (1984). *Inquiry*. Cambridge, MA: MIT Press.
- Stephens, D. W. & Krebs, J. R. (1986). *Foraging theory*. Princeton, NJ: Princeton University Press.
- Thompson, R. K. R., Oden, D. L. & Boysen, S. T. (1997). Language-naïve chimpanzees (*Pan troglodytes*) judge relations between relations in a conceptual matching-to-sample task. *Journal of Experimental Psychology: Animal Behavior Processes*, 23 (1), 31-43. [10.1037/0097-7403.23.1.31](https://doi.org/10.1037/0097-7403.23.1.31)
- Wellman, H. M., Cross, D. & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, 72 (3), 655-684. [10.1111/1467-8624.00304](https://doi.org/10.1111/1467-8624.00304)
- Wimmer, H. & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13 (1), 103-128. [10.1016/0010-0277\(83\)90004-5](https://doi.org/10.1016/0010-0277(83)90004-5)
- Wynn, K. (1992). Addition and subtraction by human infants. *Nature*, 358 (6389), 749-750. [10.1038/358749a0](https://doi.org/10.1038/358749a0)

Crediting Animals with the Ability to Think: On the Role of Language in Cognition

A Commentary on Adina Roskies

Ulrike Pompe-Alama

Davidson's argument for the claim that animals cannot be credited with beliefs rests on the assumption that possessing beliefs—as propositional attitudes—pre-supposes the possession of language. Based on Roskies' reconstruction of Davidson's argument, I want to discuss the implications of overemphasizing the role of language in thinking. I will offer a (tentative) explanation as to why this overemphasis occurs, namely due to a preoccupation with the way we experience ourselves while thinking or "having thoughts"; I further attempt to defend why a bottom-up strategy for the investigation of thought-invoking mechanisms might be a more promising way to study thought and the role of language therein.

Keywords

Beliefs | Concept of belief | Davidson | Human cognition | Language | Mental representations, | Metacognition | Non-linguistic creatures | Propositional attitudes | Thought

Commentator

[Ulrike Pompe-Alama](#)

ulrike.pompe-alama@philo.uni-stuttgart.de

Universität Stuttgart
Stuttgart, Germany

Target Author

[Adina Roskies](#)

adina.l.roskies@dartmouth.edu

Dartmouth College
Hanover, NH, U.S.A.

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

What are the defining differences between human and animal cognizers? This concern has driven philosophers and scientists for a long time,¹ well before [Darwin's \(1871\)](#) theory of evolution and its inherent claim of developmental continuity between the species. The prevailing intuition has been, and often still is, that

even though we stand in a direct developmental line with other mammals in a physiological sense, our cognitive and affective abilities far exceed theirs, not only in a quantitative, but also in a qualitative sense. Criteria to support this notion are frequently sought in an array of special cognitive abilities, such as the ability to speak (e.g., [Savage-Rumbaugh et al. 1985](#)), the

¹ See for example, Aristotle's *De anima*.

possession of concepts (e.g., Newen & Bartels 2007), or behavioral traits like altruism or cooperation (Hamann et al. 2011; Warneken 2013; Warneken & Tomasello 2009). All of these are to varying degrees attributed to humans, but are either to a much lesser degree or not at all ascribed to animals, thus representing the cornerstones of the critical divide between “us” and “them” (Hare 2007). The problem raised by Davidson and discussed by Roskies concerns the special case of beliefs and the general case of the attribution of propositional attitudes to nonlinguistic creatures.

According to Davidson, it is only in the domain of human cognition that we can sensibly apply the notion of thinking. His reasons for holding this conviction are manifold, as Roskies uncovers beautifully in her treatment of Davidson. The general line of argument will be sketched out and discussed below. Roskies refutes Davidson’s arguments mainly on empirical grounds, with the aim of establishing that nonlinguistic animals can be cognitive agents with beliefs and mental representations, which function as kinds of propositional attitudes. In this commentary, I would like to complement this line of reasoning by questioning what it takes to credit human cognizers with thoughts; or rather, what we consider to be the prerequisites for attributing thoughts and beliefs to humans. Davidson puts much weight on the possession of language. Here, I want to argue that focusing on language as a necessary cognitive instrument for being able to think poses a methodological barrier for examining what the human ability to think actually amounts to. Stressing the point that the introspectively experienced properties of thinking, a term that requires careful consideration in itself, should not be identified with and reduced to experiencing inner speech, I want to show that our understanding of what thought is needs to be complemented by a bottom-up investigation into the neural processes and mechanisms that produce higher cognitive states, such as thoughts. I argue, therefore, that our introspective access to the way thinking presents itself to us as thinkers is only one part that needs to be considered. What is required in order to understand the phenomenon of think-

ing is first a suitable conceptual framework of the notions “thought” and “thinking”, which distinguishes between their intentional and phenomenological aspects, i.e., between the content of propositional attitudes and the phenomenal states of subjects making use of these attitudes. Second, we need to show how the sub-personal and personal levels of these factors can be distinguished from each other in order to show if and how they are interconnected. These considerations will be discussed in detail after a review of Roskies’s discussion of Davidson’s account of language and belief.

2 Roskies’ reconstruction of Davidson

What Roskies dubbed Davidson’s Master Argument is a reconstruction of Davidson’s position, capturing in a nutshell both his basic assumptions about how we understand others and the background to his claims about human cognition. As Roskies puts it:

According to Davidson’s interpretationism, having beliefs entails being an interpreter. The basic idea of the Master Argument is that possessing certain concepts is a prerequisite for being an interpreter, and that an organism must have language in order to have these concepts. [...] the Master Argument links thought to language by way of higher order thoughts. Specifically, Davidson suggests that a concept of *belief* is a prerequisite for propositional attitudes, and that a concept of belief is unavailable without language. ([this collection](#), pp. 6–7)

According to Roskies, then, Davidson is forced to endorse the view that a cognizer must know what beliefs are in order to have them. Can Davidson’s view be sound? It might be correct to claim that a cognizer must possess the concept of belief to recognize *herself* as having them or to be able to attribute such a state to herself. This seems to be an act of metacognition, in which a subject scrutinizes her own mental states and recognizes them as mental states of a special kind. But is having the concept of belief necessary for first-order cognit-

ive acts, i.e., simply believing a proposition of some kind without reifying this state *as* a belief state?

Before delving into this line of thought, let us review Roskies' structural reconstruction of Davidson's Argument.

M1 If S has propositional attitudes, then S has beliefs.

M2 If S has beliefs, then S has a concept of belief.

M3 If S has a concept of belief, then S has language.

MC If S has propositional attitudes, then S has language.

M1 seems to be correct, if a belief is seen as a paradigmatic kind of propositional attitude.

M2 is a critical premise of Davidson's Master Argument, as we have already indicated above. The question in play here is: does having a belief automatically entail the possession of the concept of belief? We will discuss this point once again further below.

M3 is refuted by Roskies with the help of studies on false belief comprehension in prelinguistic infants (e.g., Onishi & Baillargeon 2005). However, a further point might be made here: M3 might indeed hold if having any concept at all implies the possession of language. However, there are models of non- and pre-linguistic concept possession (cf. Mandler 2004; Newen & Bartels 2007), which allow us to explain concept acquisition during development; theories presupposing language as necessary prerequisite for the possession of concepts, however, fail to do so.

Roskies' main criticism targets the notion of the "concept of belief". She aims to show that Davidson employs the concept of belief inconsistently throughout his argument. If this is so, then the argument fails due to equivocation.

According to Roskies, Davidson's conception of belief can be understood in three ways. She distinguishes three kinds of conceptions of belief: "on this robust view, having a concept of belief is an epistemologically-rich notion that entails having an ability to pass the 'false belief

test'" (this collection, p. 7); the so-called deflationary conception, in which "belief can come apart from reality", (ibid.) and which amounts to "having the concept of an objective reality"; and last, the so-called intermediate concept of belief, which "involves the ability to attribute representational mental states to oneself and others", (ibid.). The intermediate concept of belief, as its name implies, is intended to be a weaker notion than the robust one. In the remainder of the paper, Roskies deconstructs each reading, providing empirical examples with the aim of showing why and how Davidson fails to make his decisive point, namely, that language is a necessary prerequisite for holding beliefs.

The robust conception of belief is convincingly refuted by studies on the ability to understand counterfactual beliefs in others, as demonstrated by the so-called false belief test. Children only display the possession of a concept of belief when they pass the false belief test, usually at around the age of three to four years.² It is implausible, though, not to ascribe propositional attitudes to them (in a first-order sense) prior to having acquired such a robust notion of belief. It can even be claimed that they need the ability to ascribe propositional attitudes to develop a robust notion of belief in the first place. Thus, the robust conception of belief is not linked to having propositional attitudes and Davidson's premise M2 fails, if belief is understood in the robust sense.

The second reading of belief, the deflationary view, can be read out of Davidson's stance on so-called triangulation³ as a means of understanding objects as part of a reality external to us—via linguistic interaction with another person. However, as Roskies rightly states, the ar-

² See however, Apperly & Butterfill (2009) and Butterfill & Apperly (2013).

³ The notion of *triangulation* that appears in Davidson's later works, replacing the notion of the so-called *omniscient interpreter*, captures the idea that we can only attribute mental (propositional) attitudes to others by interpreting their utterances. In both instances, we identify contents: the content of the utterance as well as the content of the underlying mental attitude. This is, according to Davidson, a necessary unit: without an utterance, we cannot ascribe determinate propositional attitudes, which is why Davidson is committed to the view that non-linguistic creatures cannot be interpreted, at least not in a way that allows for the ascription of thoughts. This does not imply that Davidson has to negate mental states in animals, but it does mean that we cannot understand these mental states. The issue of interpretability will be raised below (see issue #3).

gumentative force of forging the link between language as a means of recognizing the external as external, making it thus objective, is quite weak. Further, it would strike us a bit of an overreach, if not as absurd, to assume that non-linguistic creatures cannot develop any sense of the external world as being external to them.

The third and final understanding of belief à la Roskies, the so-called intermediate view, stating that animals understand other animals as having mental representations of some sort, which are behaviorally relevant, rests on empirically undecided ground; here, however, the tight connection between having beliefs and possessing the concept of belief is called into question.

Having a concept of belief might be important for reflective capacities, as we want to attribute them to rational agents that must be capable of justifying their actions, but not important for having beliefs:

perhaps being a believer requires being able to think of oneself as a believer, and thus requires the concept of belief. [...] However, while there are arguments that the ability to think about oneself as a believer is required for a rich construal of theoretical rationality (see Bermúdez 2003, Ch. 7), there is no clear argument why such reflective ability should be constitutive of having beliefs. (Roskies [this collection](#), p. 11)

Roskies has thus shown that the connection between propositional attitudes and the possession of a concept of belief (and its dependence on language possession), which Davidson tries to establish, cannot be held in light of the diverging readings of the notion of the concept of belief. Thus, Davidson's strategy fails.

3 Beyond animal cognition: The case of understanding human thought

Davidson's standpoint, from which his thesis makes sense and is plausible, begins from his assumption that "radically different representation schemes" (Roskies [this collection](#), p. 2) gov-

ern in animals and humans.⁴ However, such an assumption clearly opens up a plethora of new issues. Roskies targets these by drawing attention to the empirical concerns mentioned above, thereby showing that an empirical foundation to support Davidson's background assumption is missing.

To my mind, these further issues resulting from Davidson's background assumption are the following:

1) How can we defend the intuition that animal and human cognition differ in kind? In order to defend this view, it would seem that one needs to identify a distinguishing criterion that can account for the diverging representation schemes. It also has to be shown that this factor is responsible for abilities that one group of cognizers has and that is at the same time missing in the other group. If language possession were to count as such a factor, it remains to be shown which abilities hinge on its possession and execution. At the same time, following this approach, it apparently needs to be established that no non-linguistic creature cannot execute a similar ability, not even in a partial or proto-form. This difficulty leads us to issue 2:

2) How can we understand representation schemes in animals if we do not suppose a kinship to our own cognition? As Roskies rightly states, we cannot but credit animals with numerous cognitive abilities, given their at times complex and often obviously intelligent behavior. Interpreting this behavior without acknowledging any dependence on sensory states, memory, and certain motor skills, affects, and even social competencies, seems impossible. The representation schemes employed crucially depend on physiological implementation. If the physiological basis for the acquisition of environmental information is alike in humans and

⁴ The reason why Davidson is committed to this view can be derived from the triangulation argument: since animals do not possess language, we cannot attribute determinate propositional attitudes to them. We have thus no way of knowing how they represent the world, since this is not graspable to us through our usual means of interpretation. The question, however, is whether this epistemic opacity with regard to animal cognition necessarily entails the ontological statement that their representation schema are in fact different from ours, if representation schema are seen to comprise sensory and affective states as well, and perhaps even doxastic states preceding properly expressed, i.e., propositionally coined, beliefs.

animals, how different can the representation of environmental information in terms of sensory and affective representations be? Even if a complete overlap between human and animal perception cannot be argued for on the basis of isomorphisms, we can (and perhaps must) commit ourselves to the systematicity of behavioral cause-and-effect relations. It is this systematicity that leaves little room for interpreting animal cognition (at least in the sensory and affective domain) as being radically different from ours.

3) In light of Davidson's interpretationism, how much weight does language possession carry in terms of our ability to interpret other cognitive agents? When we think of how we "make sense" of another person, we rarely rely exclusively on the other's verbal utterances. Rather, it would seem that we generally seek to compare the contents of their verbal utterances with their overt behavior; we hold another responsible, as a rational agent, if her expressed intentions diverge "too much" from her behavior. Think of the following case: your neighbor tells you about his plans to save some money for the upcoming summer vacation; the next day you see him walk into the local casino where you know he spends quite some time—and usually loses a fair amount of money. In this case, we would probably be inclined to disregard the verbal utterance ("I'm saving up for a nice summer vacation overseas"), and rather take his actions (which might involve compulsion or gambling addiction) as indicators of his real motivations and driving forces.

4) Considering this case, we can ask which role the analysis of another's beliefs play in interpreting and whether verbal utterances are a true mirror of internal thought mechanisms and proper beliefs.

5) To my mind, the most salient question is whether we can understand human cognition, especially thought, with the help of notions like beliefs (regardless of whether they are faithfully uttered or not) and their conveyance via language. Since the discussion of this issue will require some space, I shall dedicate a proper section to it below.

3.1 Experiencing oneself while thinking—the bias towards language

We can understand why Davidson (and with him many others)⁵ posits the possession of language as a necessary condition for having propositional attitudes. Namely, one may come to the view that the way a human cognizer experiences her thoughts is predominantly conveyed by her sense of inner speech.⁶ Consider for a moment what it feels like to think.⁷

Probably the most prominent, identifiable feeling related to thinking is that of your inner voice, commenting on the world around you and the world inside you, making you feel distinct from, yet embedded within it. Let's call this phenomenon—if you can follow me here—the inner-speech view⁸ with regard to thinking. I will argue that this view is misleading. Our intuitive description of what the inner-speech view comes down to is intricately linked to our ability to express the contents of our thoughts in words—the form of thoughts are, presumably, sentences that are composed of concepts and words, in our minds.

But is this identification of thought with mental speech justified? For [Vygotsky \(1934/1987\)](#), it is clear that there are large parts of thinking that do not rely on verbal expression: "There is a large range of thinking that has no direct relationship to verbal thinking" ([Vygotsky 1934/1987](#), p. 115). Such a view

⁵ In fact, my point is here not to claim that this is Davidson's motivation proper, but that we, as philosophers, can easily fall for the language-bias, language being not only the instrument but also most often the object of our trade.

⁶ One might object that Davidson's focus on language is a result of his roots in British analytic philosophy. While that is certainly true, it remains to be seen where the preoccupation with language as a "window" into the workings of the mind is derived from within this tradition; I have a hunch that the inner-speech bias I sketch plays a role here as well.

⁷ It is debated whether there is a special (that is, a unique, proprietary and distinctive) phenomenology of thinking (cf. [Bayne & Montague 2011](#)). I suspect, however, that this debate suffers from a lack of distinction between the contents (or intentional aspects) of thought and the phenomenal aspects of consciousness. The point I wish to make is that the characterization of thought we gain through introspective observation of ourselves while thinking does not grant insight into the processes that precede and produce thoughts – and this point is neutral with respect to the question whether there actually is such a thing as a distinct phenomenology of cognition.

⁸ See, for example: [Vygotsky \(1934/1978\)](#); [Watson \(1920\)](#); [Carruthers \(2002\)](#). Inner speech in Vygotsky's view means the overlap (so to speak) of our faculty of thought and our faculty of speech (cf. [Jones & Fernyhough 2006](#)).

thus allows for other, non-verbal types of thought, such as pictorial or imagistic ones, such as come to bear, for example, in mental-rotation tasks or mental imagery (Shepard & Metzler 1971; Weiskrantz 1988; Kosslyn et al. 2006).

If these instances can be found, and identified as kinds of thinking, the hypothesis that language is the one and only tool for producing thoughts in us seems simply false. That thought is exclusively verbal appears thus as a form of theory-induced illusion. One might say that the fixation on language prompted by the analytic tradition has thus resulted in the projection of the method (the analysis of language) onto the phenomenon (the human mind).

Contemporary philosophy of mind left the method of linguistic analysis behind some time ago, and in order to get away from the language-bias we should shift our focus from the surface structure of thinking, namely its intentional and phenomenological (inner-speech) characteristics, to the sub-personal level of the underlying mechanisms and production schemes of thinking.

Such a reductive approach is already in place in the numerous research efforts in cognitive science that aim at describing and explaining information processing in the brain: sensory and affective components of cognition, as well as aspects of motor behavior and memory are studied in a very promising way—in the animal as well as the human domain. The problem is that our faculty of “thinking” is in this research program a rather elusive phenomenon, for various reasons: unlike when studying the neural basis of perception, for example, thought processes cannot be studied on a cellular level, since the identification of a stimulus is virtually impossible: in vision, a stimulus is light hitting the retina, whereas the “stuff” of thought is information provided by the stimulus-processing areas, thus, an “inner-system” medium. Localizing brain areas involved in thought and thinking, on the other hand, is possible. The prefrontal cortex has been shown to be involved in planning future actions and other high-level cognitive tasks (Goldman-Rakic 1996; Fuster 2008); however, this structure is strongly con-

nected to a wide network of other cortical areas and imaging studies show that high-level cognitive tasks often if not always result from correlated activity in multiple areas across the whole of the cortex (Fuster 2008) which makes the individuation of the “center of thought” rather difficult.

In light of these complications, it is helpful to highlight the function that higher-cognitive abilities have with regard to our overall behavior. Most researchers and philosophers would agree that what this involves is the conscious representation of objects, including the deliberate manipulation of information, retrieved from memory as well as from present and actual stimuli, for the purpose of problem solving, decision making, social interaction, communication, and action planning. The involvement of language-processing areas in the execution of these tasks has already been shown (see e.g., Goel et al. 2000)—but does this suffice to support the claim that language is a necessary cornerstone of the neural basis of higher-level cognition in humans?

When “thinking” is divided and described in terms of its functional rather than phenomenal properties, the question of how far thinking relies on our capacity to speak or use language can be replaced by the question of which brain areas and input-output relations we find involved in the faculties mentioned above. This program requires a reorientation in terms of research methods and a redefinition of the phenomenon: the phenomenological description of “thinking”, e.g., in terms of inner speech, does not supply us with an understanding of its underlying processes and mechanisms. It is these, however, that we should know first before we can put our finger on the role that language (the inner and external version alike) plays in the execution and the production of the cognitive capacities listed above.

When we cannot help but attribute the ability to manipulate information in a creative way to animals and intuitively call this “thinking” (think of the Kea, a species of bird known for its curiosity and astonishing abilities in handling difficult mechanisms—they can virtually break into a safe; cf. Auersperg et al. 2009;

Huber & Gajdon 2006; Werdenich & Huber 2006) we seem to have found a satisfactory criterion for crediting animals with a form of demanding cognition, not unlike our own, even though we cannot claim to understand what it feels like or how the world represent itself to the Kea.

Such a language-independent form of high-level cognition might rule in us as well, such that it precedes the formation of beliefs we form on states of the world and their linguistic representation. It might be the case, and this is the point I want to stress in this commentary, that we fall in a systematic way for a fallacy of experienced thinking, which presents us with a linguistic representation of the contents of thought, whereas the mechanisms producing these thoughts may not rely and are not caused by speech and language involving neural mechanisms.

One can object that this is not what Davidson had in mind when he claimed that thought depends on language. Davidson's idea rests (so goes the defense) upon the assumption that language is a universal format of information processing unique to humans (in the first place) and an instance of cognition, which lies at the core of human cognition, regardless of its temporal and causal involvement in the production of thoughts. But this—so I want to claim—amounts to a phenomenological argument, even if Davidson presents it as a theoretical one. So even if language were the universal format of human thought, the empirical basis for such a claim would be quite opaque, and any theoretical argument so far rests on this weak empirical basis.

4 Conclusion

The question of whether thought is exclusively verbal or linked to language capacity is not answerable from a phenomenological point of view, since we can think of instances of mental symbol-use that do not rely on language; on the contrary, we know that language “fills in the void”, so to speak: when we acquire language, it fulfills the cognitive demands to express references and relations among them. In this view,

thought and thinking precede the linguistic representation of the involved concepts.

If one wants to follow this line of thought, it remains to be shown how the Davidsonian *dictum* that animals do not have a special form of cognitive ability, namely, propositional attitudes such as beliefs, desires etc., relates to the general argument on higher-cognitive faculties, which do not depend on language possession and which are of the same kind across the animal and human realms. It would thus have to be argued for a language-independent form of propositional attitudes.

Does the inner-speech bias bear not only on thinking at large but also our self-attribution of desires and beliefs? It might. Roskies rightly raises the question, contra Davidson, of whether all our beliefs have definite content ([this collection](#), p. 6). In my view, as soon as we hold a belief *qua* belief, some kind of cognitive meta-representation must come into play. Such a form of meta-representation strikes me as probably being conveyed by the inner speech mechanism and as thus being subject to the phenomenological inner-speech fallacy.

Roskies nicely disassembles Davidson's arguments and reconstructs them in a clear and easy-to-follow fashion. She exposes their argumentative weaknesses (such as the issue of interpretation and behavior) and provides ample empirical examples of, and conceptual arguments for, why we should not follow Davidson in his assessment of animals' cognitive abilities. However, I have tried to show that a further underlying claim can be made, namely that not only is animal cognition a matter of speculation, but that even our own inner workings are less transparent than we commonly like to assume. Davidson's claim rests, to my mind, on the rashly embraced yet unfounded assumption that language plays a key role in higher cognition in humans (1984, 2001). In my view, contemporary research efforts in the cognitive sciences, but also in philosophy, undermines—or at least calls into question—this assumption. Certainly we are dealing with an important philosophical claim, which could only be properly backed up by extensive empirical evidence pointing to the ubiquitous involvement of language-processing

brain areas and mechanisms in higher-level cognitive tasks such as decision-making, action planning, deliberation, etc. Doubtless, human cognition benefits from the linguistic format; abstract thoughts about, e.g., liberty can probably only be executed at a significantly deep level if the relevant concepts have been provided by a linguistic community. But the need to express a certain feeling, like freedom as the opposite of (the feeling of) constraint, for example, certainly originates in a pre-verbal or non-verbal manifestation of this feeling.

Focusing on language, therefore, blocks a fuller examination of what thinking in humans amounts to. We have, I believe, misled ourselves in the face of the phenomenology of inner speech as to what it is like to think, for us as humans. But this gets us only part way towards a full understanding of the underlying mechanisms, structures, and sources of thoughts.

References

- Apperly, I. A. & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, 116 (4), 953-970. [10.1037/a0016923](https://doi.org/10.1037/a0016923)
- Auersperg, A. M. I., Gajdon, G. K. & Huber, L. (2009). Kea (*Nestor notabilis*) consider spatial relationships between objects in the support problem. *Biology Letters*, 5 (4), 455-458. [10.1098/rsbl.2009.0114](https://doi.org/10.1098/rsbl.2009.0114)
- Bayne, T. & Montague, M. (Eds.) (2011). *Cognitive phenomenology*. Oxford, UK: Oxford University Press.
- Butterfill, S. A. & Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind and Language*, 28 (5), 606-637. [10.1111/mila.12036](https://doi.org/10.1111/mila.12036)
- Carruthers, P. (2002). The cognitive functions of language. *Behavioral and Brain Sciences*, 25 (6), 657-674. [10.1017/S0140525X02000122](https://doi.org/10.1017/S0140525X02000122)
- Darwin, C. (1871). *The descent of man, and selection in relation to sex*. London, UK: John Murray.
- Davidson, D. (1984). *Inquiries into truth and interpretation*. Oxford, UK: Clarendon.
- (2001). *The subjective, intersubjective, objective*. Oxford, UK: Clarendon.
- Fuster, Joaquin M. (2008). *The prefrontal cortex*. Boston, MA: Academic Press.
- Goel, V., Buchelt, C., Frith, C. & Dolan, R. J. (2000). Dissociation of mechanisms underlying syllogistic reasoning. *NeuroImage*, 12 (5), 504-514. [10.1006/nimg.2000.0636](https://doi.org/10.1006/nimg.2000.0636)
- Goldman-Rakic, P. S. (1996). The prefrontal landscape: Implications of functional architecture for understanding human mentation and the central executive. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 351 (1346), 1445-1453. [10.1098/rstb.1996.0129](https://doi.org/10.1098/rstb.1996.0129)
- Hamann, K., Warneken, F., Greenberg, J. A. & Tomasello, M. (2011). Collaboration encourages equal sharing in children, but not in chimpanzees. *Nature*, 476 (7360), 328-331. [10.1038/nature10278](https://doi.org/10.1038/nature10278)
- Hare, B. (2007). From nonhuman to human mind: What changed and why? *Current Directions in Psychological Science*, 16 (2), 60-64. [10.1111/j.1467-8721.2007.00476.x](https://doi.org/10.1111/j.1467-8721.2007.00476.x)
- Huber, L. & Gajdon, G. K. (2006). Technical intelligence in animals: The kea model. *Animal Cognition*, 9 (4), 295-305. [10.1007/s10071-006-0033-8](https://doi.org/10.1007/s10071-006-0033-8)
- Jones, S. R. & Fernyhough, C. (2006). Neural correlates of inner speech and auditory verbal hallucinations: A critical review and theoretical integration. *Clinical Psychology Report*, 27, 140-154. [10.1016/j.cpr.2006.10.001](https://doi.org/10.1016/j.cpr.2006.10.001)

- Kosslyn, S. M., Thompson, W. L. & Ganis, G. (2006). *The case for mental imagery*. Oxford, UK: Oxford University Press.
- Mandler, J. M. (2004). *The foundations of mind: The origins of conceptual thought*. New York, NY: Oxford University Press.
- Newen, A. & Bartels, A. (2007). Animal minds and the possession of concepts. *Philosophical Psychology*, 20 (3), 283-308. [10.1080/09515080701358096](https://doi.org/10.1080/09515080701358096)
- Onishi, K. H. & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, 308 (5719), 255-258. [10.1126/science.1107621](https://doi.org/10.1126/science.1107621)
- Roskies, A. (2015). Davidson on believers: Can nonlinguistic creatures have propositional attitudes? In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Savage-Rumbaugh, S., Rumbaugh, D. M. & McDonald, K. (1985). Language learning in two species of apes. *Neuroscience and Biobehavioral Reviews*, 9 (4), 653-665. [10.1016/0149-7634\(85\)90012-0](https://doi.org/10.1016/0149-7634(85)90012-0)
- Shepard, R. & Metzler, J. (1971). Mental rotation of three dimensional objects. *Science*, 171 (3972), 701-703.
- Vygotsky, L. S. (Ed.) (1987). *Thinking and speech. The collected works of L. S. Vygotsky, vol. 1*. New York, NY: Plenum.
- Warneken, F. (2013). The development of altruistic behavior: Helping in children and chimpanzees. *Social Research*, 80 (2), 431-442. [10.1353/sor.2013.0033](https://doi.org/10.1353/sor.2013.0033)
- Warneken, F. & Tomasello, M. (2009). Varieties of altruism in children and chimpanzees. *Trends in Cognitive Science*, 13 (9), 397-402. [10.1016/j.tics.2009.06.008](https://doi.org/10.1016/j.tics.2009.06.008)
- Watson, J. B. (1920). Is thinking merely the action of language mechanism? *British Journal of Psychology*, 11 (1), 87-104. [10.1111/j.2044-8295.1920.tb00010.x](https://doi.org/10.1111/j.2044-8295.1920.tb00010.x)
- Weiskrantz, L. (1988). *Thought without language*. Oxford, UK: Oxford University Press.
- Werdenich, D. & Huber, L. (2006). A case of quick problem solving in birds: String pulling in keas, *Nestor notabilis*. *Animal Behaviour*, 71, 855-863. [10.1016/j.anbehav.2005.06.018](https://doi.org/10.1016/j.anbehav.2005.06.018)

Thought, Language, and Inner Speech

A Reply to Ulrike Pompe-Alama

Adina Roskies

Pompe-Alama's commentary raises interesting issues regarding the nature of thought and its relation to language. She underlines the evolutionary relationship we have to other animals and results from cognitive science to argue that human thought is probably not fundamentally linguistic, and notes that the pull of the phenomenal experience of inner speech may mislead us into thinking it is. While I agree with these claims, I disagree that Davidson's own arguments are predicated on an inner speech view, and raise problems for the idea that functional imaging will easily resolve the debate about the relation of thought and language.

Keywords

fMRI | Inner speech | Language | Propositional attitudes | Representation

Author

[Adina Roskies](#)

adina.l.roskies@dartmouth.edu

Dartmouth College
Hanover, NH, U.S.A.

Commentator

[Ulrike Pompe-Alama](#)

ulrike.pompe-alama@philo.unistuttgart.de

Universität Stuttgart
Stuttgart, Germany

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

I largely concur with Pompe-Alama's commentary on my contribution to this collection. She nicely summarizes my arguments against what I call Davidson's "Master Argument," an argument that he levies against the possibility of propositional attitudes for nonlinguistic animals. As Pompe-Alama notes, aside from conceptual clarifications, my arguments are largely empirical. As such, the strength of my arguments depends on the solidity of the empirical facts they are based upon. But provisionally, since all the logically valid reconstructions of Davidson's arguments have what look to be em-

pirically false premises, none serves to establish the impossibility of animal thought.

Pompe-Alama then offers an interesting discussion of the Davidsonian claim that nonlinguistic animals cannot have propositional attitudes. She locates the source of the dispute at the phenomenological level, citing the phenomenology of thought as "inner speech", and suggests that it is this that leads Davidson, and us, to mistakenly think that thinking is fundamentally a language-dependent phenomenon. While I disagree that this is the source of Davidson's perspective, I appreciate Pompe-Alama's

discussion of some important practical consequences of the Davidsonian view, or any view that posits human thought processes to be qualitatively different than those of all other animals. In her discussion, Pompe-Alama tells us that contemporary cognitive science indicates that Davidson is wrong, and suggests that our own understanding of our own thought processes may be adversely influenced by our introspective recognition of our thoughts as embodied in inner speech. She cautions that too much attention to the phenomenological or introspective sense of inner speech can prevent us from exploring the representational aspects and physiological bases of thought that we share with other animals, and moreover, she suggests that taking language to be a necessary prerequisite for thinking poses a barrier to understanding human thought as well. As a remedy, she suggests that we discount the phenomenal aspects of thinking and instead focus on a reductive strategy for exploring the neural basis of human and animal thought in a bottom-up fashion.

2 Inner speech

Pompe-Alama calls attention to the “feeling of what it is like to think”, which she identifies as the experience of our thoughts as inner speech. There is of course debate about whether it feels like anything at all to think. However, regardless of whether our recognition of inner speech is a feeling or a cognitive introspective conclusion, this phenomenon certainly plays a role in the general tendency to and perhaps our willingness to identify thought with language. But Pompe-Alama’s easy identification of the phenomenology of inner speech with Davidson’s denial of animal thought threatens to trivialize what I take to be a fairly sophisticated, if incorrect, view about the nature of animal thought. Davidson’s interpretationism is the root of his denial, and his target is specifically propositional thoughts and related attitudes, not cognitive processing more generally. Pompe-Alama cites Vygotsky’s claim that lots of thought is not verbal thought, and she suggests that pictorial or imagistic thought should be possible

for non-linguistic creatures. I don’t suppose Davidson would refuse to recognize that animals have complex representations and even some relatively high-level cognitive capacities. But he would deny that these forms of thought had propositional contents. So the real question at issue is whether the representational power afforded by representations in nonlinguistic animals allows them to represent propositions.

That said, Pompe-Alama’s claim that the restriction of thought to verbal vehicles may be a “theory-induced illusion” is well taken. The tendency to think that only language-like formulations allow propositional content to be captured or delineated seems ungrounded, especially since philosophy has supplied us with non-linguistic means of representing propositions (Stalnaker 1987), or alternatives to propositional attitudes (Churchland 1992). Undoubtedly, propositional content requires some kind of framework that permits complex structural relationships between representations, but there is no a priori reason to think that such structure can only be achieved with linguistic implementation. Pompe-Alama is correct to point out that in our own interpretation of others, we often privilege behavior over self-report, and much social science has suggested that words, and indeed even one’s own introspective thoughts, are not a reliable window into higher cognitive processes. She also mentions that our own interpretational skills, applied to animals, yields attributions of cognitive processes that are in many ways akin to our own. Indeed, we easily attribute to them propositional attitudes. These observations put pressure on Davidson’s view, and raise the question of what our own propositional attitudes may endow us with, cognitively speaking, that the presumptively propositional-attitudeless animals are missing, if in fact he turns out to be right.

Pompe-Alama doubts whether language really plays a key role in human higher cognitive functions. We know it certainly does in one of them: Linguistic cognition. Whether it plays a fundamental role in other aspects of higher cognition is yet unknown. Davidson himself is not clear about whether he thinks language is necessary as a vehicle for thought. This distin-

guishes him from Fodor, who also thinks language is central to thought, but posits a mental language to serve as the vehicle of thought, and that is available to linguistic and non-linguistic creatures alike. Davidson's view is more subtle, and seems to depend more on social/interpersonal factors and abilities or dispositions than on vehicles per se. Thus, for Davidson, the fact that we can identify instances of non-linguistic symbol use in high-level thought is not telling, since it is the fact that we are language-using creatures that is of prime importance. It is within Davidson's purview to claim that our mastery of language makes possible thoughts that rely on non-linguistic (yet symbolic) properties.

3 Methodological difficulties

Pompe-Alama suggests that to lessen the grip of the illusion, we must pay attention to the low level realization of our thoughts. That is of course a goal of many cognitive neuroscientists, but as Pompe-Alama well recognizes, it is a difficult one to achieve. Unlike perception and action, both which can be correlated with measurable external phenomena (perception with the stimuli occurring in the external world; action with elicited motor activity), thoughts are seemingly spontaneous, and largely uncoupled from immediate environmental stimulation and control. The unpredictability of the content and occurrence of our thoughts, together with the fact that we have no idea how they are realized in neural activity (and thus which aspects of the remarkably complex signals we can record from the brain are relevant), has the consequence that thoughts promise to be extremely difficult to measure scientifically. What exactly are we supposed to look for in signals from neural tissue that is supposed to correspond to propositional thoughts as opposed to other (non-propositional) forms of mental representation? Unless we discover some means of answering this question, it will be difficult to determine empirically whether other animals have the capacity for propositional thought or not.

Taking a reductive approach, [Pompe-Alama](#) says "the question of how far thinking

relies on our capacity to speak or use language can be replaced by the question of which brain areas and input-output relations we find involved in the faculties mentioned above" ([this collection](#), p. 6). She suggests that the progress we have made in understanding the neural basis of language processing could help us resolve the debate about whether human and nonhuman cognitive processes are fundamentally different. Work in cognitive science has shown that a network of brain areas seem consistently linked with processing of natural language. Pompe-Alama suggests that we could approach the question of whether human thought is primarily linguistic by determining with functional imaging whether these areas are consistently active during human propositional thought. This will not be determinative, for reasons I sketch here. Most importantly, even if we do see activity in these areas, it will not serve to answer the question of whether human thought is fundamentally linguistically-based. Suppose phenomenal inner speech typically accompanies our thought, and it is dependent on activity in these areas. This may be because our thoughts are fundamentally linguistic, but it could also be merely a causal consequence of the deeper thought processes, without constituting them or being a necessary component of them at all. Thus, if we consistently saw activity in language-relevant areas, it might not be reflective of the fundamental nature of our thought. Suppose, on the other hand, that we failed to see such activation (and suppose we knew that inner speech was dependent on activation of language areas). This could be due to the low signal-to-noise ratio of the methods, or to the fact that language pervades brain representation and is not restricted to the areas that we typically see "light-up" in a language task, or it could indicate the non-linguistic nature of thought. In this domain, negative results are not decisive. Thus, the question of whether language centers are always active during human propositional thought will not resolve the issue.

That said, significant progress is being made in understanding at least some aspects of the representational coding of thought contents. The object perception literature demonstrates

that cognitive neuroscience has achieved much in the last few years, due to work with both noninvasive fMRI in humans and invasive recording in humans and nonhuman primates. In particular, we have gained much greater insight into the representational coding of faces, with access to regional information about coding of representational aspects of face identity, similarity, expression, and so on (see e.g., [Haxby et al. 2014](#), and [Freiwald & Tsao 2011](#)). Other work suggests that the visual cortex represents semantic features in the form of a cortical map ([Huth et al. 2012](#)). Although this kind of work is in its infancy, novel analytical and modeling techniques promise to continue to yield a deeper understanding of how our brains represent semantic properties. An important result stemming from this kind of research is evidence of the extensive homologies between neural processes of visual representation in humans and nonhuman primates ([Sha et al. in press](#); [Kiani et al. 2007](#)). These homologies seem to extend in large part to complex cognitive processes such as decision-making ([Gold & Shadlen 2007](#)). At the neural level, we have no evidence of qualitative differences in neurological processing between humans and nonhuman primates, nor evidence that we and they possess radically different representational frameworks. Nonetheless, none of the work mentioned explicitly targets propositional contents, and very little extant work has looked at the combinatorial or structural properties of these mental representations. In my own view, answers to these difficult questions will not come from bottom-up approaches alone or even in large part. Only a high-level theory of brain function is likely to make real headway on this issue. It will be interesting to see whether new work in predictive coding (see [Clark this collection](#); [Hohwy this collection](#); [Seth this collection](#)) allows for new ways of approaching these fundamental questions.

4 Conclusion

Pompe-Alama seems to argue that Davidson's argument about the impossibility of animal thought is at base an argument based on the phenomenology of thought as inner-speech. I

don't see this. His is an argument about the process of interpretation, and the interpersonal nature of objective thought. While I disagree with Davidson's arguments, and in particular with the view that animals cannot have propositional attitudes, I am nonetheless sympathetic to the possibility that the ability to use language makes possible cognitive feats that are unavailable to nonlinguistic creatures (see e.g., [Roskies 2015](#)). These may only be quantitative differences, allowing us to represent contents that nonlinguistic creatures cannot represent, or they may be more qualitative leaps, such as giving us metarepresentational abilities that make possible culture, cross-generational learning, and science. Thus, whether Davidson is right or wrong, we are still left with the fascinating question: What does language or linguistic competence allow us to do that we otherwise couldn't do?

References

- Churchland, P. (1992). *A neurocomputational perspective: The nature of mind and the structure of science*. Cambridge, MA: MIT Press.
- Clark, A. (2015). Embodied Prediction. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1-21). Frankfurt a.M., GER: MIND Group.
- Freiwald, W. R. & Tsao, D. (2011). Taking apart the neural machinery of face processing. In A. J. Calder, G. Rhodes, M. H. Johnson & J. V. Haxby (Eds.) *Handbook of face perception* (pp. 707-718). Oxford, UK: Oxford University Press.
- Gold, J. I. & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, 30, 535-574. [10.1523/JNEUROSCI.1939-07.2007](https://doi.org/10.1523/JNEUROSCI.1939-07.2007)
- Haxby, J. V., Connolly, A. C. & Swaroop Guntupalli, J. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annual Review of Neuroscience*, 37, 435-456. [10.1146/annurev-neuro-062012-170325](https://doi.org/10.1146/annurev-neuro-062012-170325)
- Hohwy, J. (2015). The neural organ explains the mind. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1-22). Frankfurt a.M., GER: MIND Group.
- Huth, A. G., Nishimoto, S., Vu, A. T. & Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the brain. *Neuron*, 76 (6), 1210-1224. [10.1016/j.neuron.2012.10.014](https://doi.org/10.1016/j.neuron.2012.10.014)
- Kiani, R., Esteky, H., Mirpour, K. & Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology*, 97 (6), 4296-4309. [10.1152/jn.00024.2007](https://doi.org/10.1152/jn.00024.2007)
- Pompe-Alama, U. (2015). Crediting Animals with the Ability to Think: On the Role of Language in Cognition—A Commentary on Adina Roskies. *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Roskies, A. L. (2015). Monkey decision making as a model system for human decision making. In A. Mele (Ed.) *Surrounding free will* (pp. 231-254). New York, NY: Oxford University Press.
- Seth, A. K. (2015). The cybernetic Bayesian brain. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1-25). Frankfurt a.M., GER: MIND Group.
- Sha, L., Haxby, J. V., Abdi, H., Guntupalli, J. S., Oosterhof, N. N., Halchenko, Y. O. & Connolly, A. C. (in press). The animacy continuum in the human ventral pathway. *Journal of Cognitive Neuroscience*
- Stalnaker, R. C. (1987). *Inquiry*. Cambridge, MA: MIT Press.

Bridging the Objective/Subjective Divide

Towards a Meta-Perspective of Science and Experience

Jonathan Schooler

In this paper I use the thesis that perspective shifting can fundamentally alter how we evaluate evidence as the backdrop for exploring the perennial challenge of bridging the divide between the subjective first-person perspective of experience, and the objective third-person perspective of science. I begin by suggesting that reversible images provide a metaphor for conceptualizing how the very same situation can be understood from two very different perspectives that appear to produce seemingly irreconcilable accounts of their contents. However, when one recognizes that both views are different vantages on some deeper structure, a meta-perspective can emerge that potentially offers a vantage by which the opposing perspectives can be reconciled. Building on this notion of a meta-perspective, I outline a framework for conceptualizing how science can draw on individuals' first-person experience in order to explicate those experiences within the necessarily third-person perspective of science. I then show how this approach can illuminate one of the most private yet ubiquitous aspects of mental life: mind-wandering. Finally and most speculatively, I attempt to tackle the enduring ontological tensions that emerge from the disparities between the first- versus third-person perspectives. Specifically, I suggest that the present prevailing third-person perspective of material reductionism fails to adequately account for the first-person experience of subjectivity, the flow of time, and the present. While I argue that these differences are an intrinsic property of each perspective, and thus irreconcilable from the vantage of either, I raise the possibility of a meta-perspective in which these clashes might be better accommodated. Toward this end, I speculatively suggest that experience, the flow of time, and the unique quality of "now" might be accommodated by the postulation of a subjective dimension or dimensions of time.

Keywords

Consciousness | Heterophenomenology | Meta-awareness | Meta-perspective | Mind wandering | Mind/body problem | Neurophenomenology | Neutral monism | Panpsychism | Phenomenology | Time

1 Introduction

I am the proud owner of a philosopher's stone. Although it does not hold any of the mysterious powers (e.g., turning lead to gold, providing endless youth) that the alchemists attributed to its namesake, I nevertheless feel its title fitting, as it offers some rather deep insights into the importance of perspective in defining what seems true. What distinguishes my stone from

an ordinary river rock is that it has engraved upon it the statement "Nothing is written in stone." In pondering its irony, I've come to realize that my philosopher's stone can be viewed in at least three ways, each leading to a different accounting of its merit. From one vantage the statement on the stone is self-evidently false, as clearly revealed by where it is carved.

Author

[Jonathan Schooler](#)
jonathan.schooler@psych.ucsb.edu
University of California
Santa Barbara, California, United States of America

Commentator

[Verena Gottschling](#)
vgott@yorku.ca
York University
Toronto, Ontario, Canada

Editors

[Thomas Metzinger](#)
metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

From another it is demonstrably true, as the word “nothing” is written in stone right there. Finally, the fact that the presentation of the stone’s message simultaneously reveals it to be both true and not true enables the stone to clarify the paradoxical essence of its meaning. Nothing is definitive because a change in perspective may shift what is seen as factual. However, the stone further illustrates that when one recognizes how the perspectives that one takes influence the conclusions that one draws, one gains a larger meta-perspective that can accommodate them both.

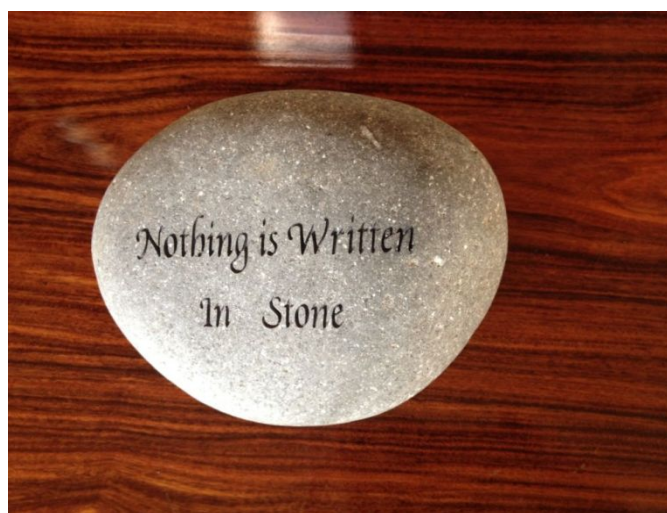


Figure 1: From one perspective, as evident from the place onto which it is carved, “Nothing is Written in Stone” is a contradictory statement. However, from another perspective, “Nothing” is in fact written in stone, making the statement true. Thus, “Nothing is Written in Stone” illustrates that when one recognizes how the perspectives that one takes influence the conclusions that one draws, one gains a larger meta-perspective that can accommodate them both.

Although it is relatively straightforward to describe the manner in which my philosopher’s stone conveys how shifting perspective can alter what is seen as true, such descriptions do not do justice to the impact the stone has when one actually encounters it. The stone not only conveys its message, it embodies it. Its message thus speaks not only to one’s capacities of logic but also viscerally, physically, through one’s senses. Indeed this difference between the third-person *account* of something and the first-person *experience* of it is perhaps the ultimate ex-

ample of the manner in which perspective can alter how we understand the world.

In this paper I attempt to nudge the field towards a rapprochement between the subjective first-person perspective of experience and the objective third-person perspective of science. My efforts are divided into three somewhat distinct sections; all united by the goal of illustrating how the divide between the subjective and objective might begin to be bridged by a broader perspective that acknowledges that while neither can be reduced to the other, they may be alternative vantages of a larger meta-perspective.

In the [first](#) section, I use the analogy of reversible images to emphasize the importance of perspective shifting in recognizing that views that seem one way from one perspective may seem quite different from another. However, when one recognizes that both views are different vantages on some deeper structure, a meta-perspective can emerge that potentially offers a vantage by which the opposing perspectives can be reconciled. I propose that the relationship between the first-person perspective of subjective experience and the third-person perspective of objective science can be conceptualized in this manner, and that at least some of the heated debate between scholars on this topic may stem from their exclusively favoring one vantage over the other.

In the [second](#) section, I illustrate how the third-person perspective of science can both draw on and elucidate first-person experiences, and in particular the ubiquitous internal state of mind-wandering. I argue that although people’s self-reports of private internal experiences such as mind-wandering necessarily rely on a re-representation of the experience to themselves (meta-awareness), we can nevertheless draw inferences about their underlying experience by examining the relationship between self-reports and physiological and behavioral measures. Triangulation between these measures has highlighted both the strengths and limitations of people’s meta-awareness of their drifting minds: although people frequently fail to notice that their minds are wandering, when queried they are quite accurate at reporting whether

or not their minds were on task. This analysis thus reveals the value of using empirical third-person science to clarify the nature of first-person experience.

In the [final](#) section I consider how first-person experience may inform our understanding of objective reality. Current views of science offer no way of accounting for the existence of subjective experience, the flow of time, or the privileged present, leading mainstream science to marginalize these essential elements of consciousness as irrelevant or illusory. However, from my vantage these aspects of existence are at least as certain as physical reality itself. It seems nonsensical to characterize experience as an illusion, because even an illusory experience (i.e., where the contents have no bearing on physical reality) is still an experience. Moreover, experience exclusively resides in an ever-changing present. A characterization of reality that has no place for subjective experience, the flow of time, or importance of the present seems devoid of the core aspects of my existence. In keeping with others who have speculated that theories of physical reality will need to be expanded to accommodate subjective experience, I conjecture that consciousness may correspond to movement in an additional subjective dimension (or dimensions) of time. Although this hypothesis is highly speculative, it provides an example of the kind of meta-perspective that may be necessary to successfully accommodate subjective and objective views.

Clearly I have my work cut out for me. However, before embarking on the more ambitious aspects of this journey, let us first step back and consider the nature of perspective and the impact that it can have on understanding.

2 Applying perspective shifts to conceptualizing human experience from the first- versus third-person perspective

The striking parallels between perceptual and conceptual perspective shifts exemplify the embodiment of mental capacities in physical experience ([Schubert & Semin 2009](#)). Colloquially, when we talk about dramatic shifts in concep-

tual understanding, we routinely use perceptual metaphors ([Schooler et al. 1994](#)). We speak of “thinking out of the box,” or of “stepping back and looking at the bigger picture.” Even the term that we use for gaining a fresh perspective on an old problem, i.e., “insight,” directly alludes to the parallels between perceptual and conceptual perspective shifting. It is no coincidence that the Gestalt psychologists who pioneered research on visual perspective shifting ([Wagemans et al. 2012](#)) also were the first to investigate the processes of conceptual insight ([Duncker 1945](#)). And indeed, research in our lab ([Schooler & Melcher 1995](#)) reveals a strong correlation between people’s ability to make perceptual insights (e.g., recognizing out-of-focus pictures) and conceptual insights (e.g., solving insight word problems). Thus, in order to explore how perspective may constrain our conceptual understandings, it is helpful to start by briefly considering the ways in which perspective can influence perceptual experiences. As will be argued, the manner in which alternative first-person perceptual perspectives constrain our experiences, provides a compelling metaphor for the broader contrast between first- and third-person perspectives that individuals face in reconciling their personal subjective experiences with objective reality.

One of the greatest challenges of visual perspective is recognizing how fluid it really is. Typically, when we view an object or a scene, we apprehend it from a particular vantage and rarely consider the possibility that it may be seen in a different way. If and when a shift occurs, the experience is typically characterized by a marked surprise that the very same view could afford such a different understanding. The Gestalt reversible figures are a quintessential example of images that startle us with their alternative perspectives. At first when we encounter them we often perceive them from only one perspective; that is, although there are several possible interpretations of the image, we assign one set of perceptual properties to the elements of the image (front or back, figure or ground), and one conceptual interpretation of the object (e.g., duck or rabbit, young woman or old hag).

When presented with an image of a duck/rabbit as a duck, those unfamiliar with the image may initially see only a duck. However, if alerted to the possibility of another embedded image, suddenly a rabbit may virtually pop out. Other classic examples of reversible images include: a Necker cube facing one way or another, a vase or a pair of faces, a young woman or an old hag. A particularly compelling recent addition is the spinning dancer illusion, where a perceptual shift not only changes one's perspective of her orientation but also the direction in which she appears to be spinning.



Figure 2: The duck-rabbit illusion is a classic example of a perspective-dependent reversible image. When presented as a duck, those unfamiliar with the image may initially see only a duck. However, if alerted to the possibility of another embedded image, suddenly a rabbit may pop out. McManus, I. C., Freegard, M., Moore, J., & Rawles, R. (2010). Science in the making: Right hand, left hand. II: The duck-rabbit figure. *Laterality*, 15, 167.

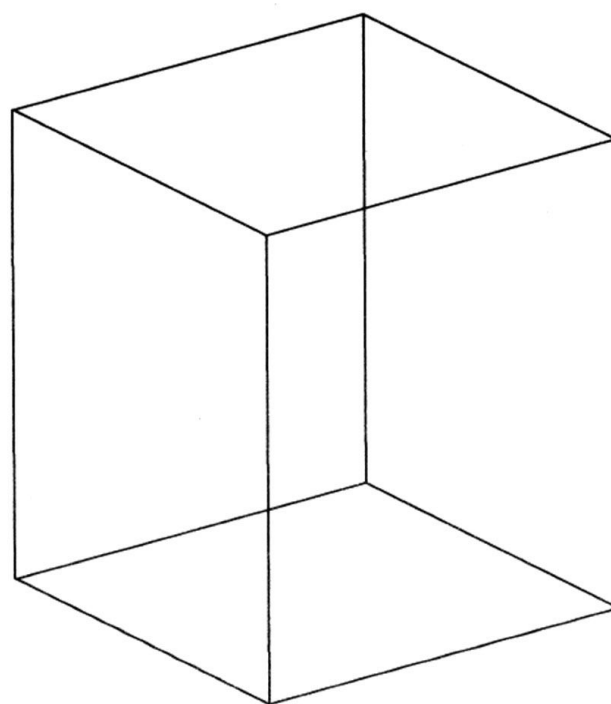


Figure 3: The Necker cube is a reversible image that, depending on the perspective taken by the observer, appears to be facing one way or another. Shifting one's perspective allows the observer to view the cube either from slightly above or slightly below. Necker, L.A. (1832). Observations on some remarkable optical phenomenon seen in Switzerland; and on an optical phenomenon which occurs on viewing a figure of a crystal or geometrical solid. *London and Edinburgh Philosophical Magazine and Journal of Science*, 1 (5), 329–337.

There are several notable aspects of all the aforementioned visual perspective shifting examples. First, before one knows that there are multiple interpretations, it is common to only perceive one or the other. Second, once one is aware of both perspectives, one can experience an oscillation between the two, shifting from one perspective to the other, and back again. Third, at any one moment in time, it is impossible to simultaneously see both interpretations. The Necker cube is either seen facing one way or the other; the spinning dancer only rotates in one direction at a time. Finally, although one can only perceive one interpretation at a time, one can nevertheless know that multiple perspectives exist, and this knowledge provides a *meta-perspective*, whereby we appreciate that what we

perceive one way at one moment can be perceived in a very different way in the next.

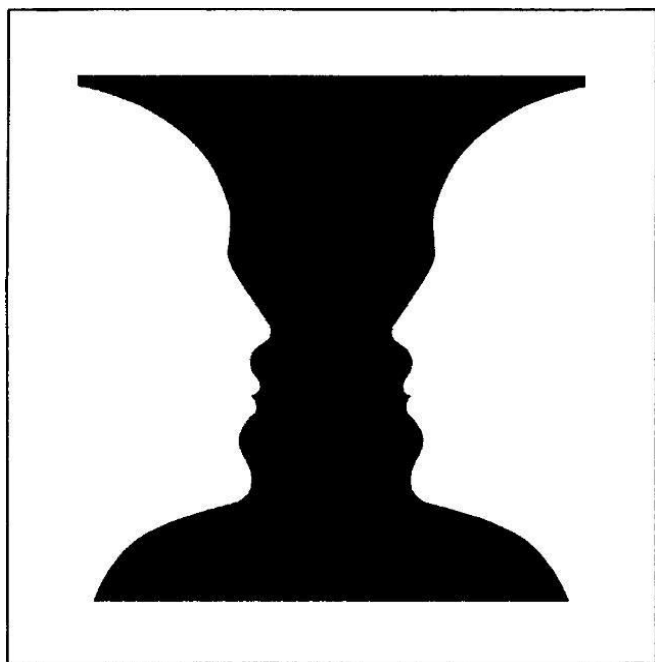


Figure 4: Rubin's vase (sometimes referred to as "The Two Face, One Vase Illusion") depicts the silhouette of a vase in black and the profiles of two inward-looking faces in white. The figure-ground distinction made by the brain during visual perception determines which image is seen. Ittelson, W. H. (1969). *Visual Space Perception*, Springer Publishing Company, LOCCCN 60-15818

A particularly remarkable class of perceptual shift that enables us to switch to a meta-perspective comes from "Magic Eye" stereograms that can reveal a full holographic three-dimensional realm that is not initially perceptible at all. These stereograms entail images that first are viewed as a two-dimensional pattern. However, if one stares at the image long enough in just the right way (this requires a little eye crossing) and believes that it is possible to actually see into it, an entirely different and fully three-dimensional image emerges. What is so striking about these "Magic Eye" stereograms is that the embedded three-dimensional images have absolutely no resemblance to the two-dimensional images from which they emerge. There is of course a sophisticated algorithm (based on principles of stereopsis) that enables the three-dimensional perception to arise from the two-dimensional

image, but the experiences of the two images are wholly of a different sort. Those who have not gotten into a Magic Eye image can have no idea what the underlying image looks like, and even if they are shown what the form is, they cannot appreciate what it is like to actually witness the two-dimensional page miraculously open up into a three-dimensional world that is somehow residing within it. However, those who have experienced this transformation gain a wholly different appreciation for the image, recognizing that it affords two entirely different vantages, even while appreciating that only one can be apprehended at any particular time.¹

The lessons learned from perceptual perspective shifting are relevant to the long-standing tension between conceptualizing human experience from the first- versus third-person perspective. Not unlike the shifting perspectives of a reversible image, the field of psychology has vacillated back and forth between focusing on people's self-reported internal experiences (the first-person perspective) and their observable behaviors (the third-person perspective). Moreover, just as the spinning dancer can move in one direction for a while, then flip back and forth in direction, and then carry on in the opposite direction, the field has had periods of relative steady focus on one or the other vantage and other periods in which the vantage was more variable.

¹ A possible objection to the Magic Eye stereogram as an illustration of a shifting perspective is that it can be enabled merely by a musculature action (the crossing of the eyes). One reviewer suggested that the new representation that emerges from these images may be no "more interesting than the muscular action of opening a closed eye which also allows the appearance of a suddenly unseen picture." While a worthwhile observation, I do not think it challenges the relevance of the example. First, closing one's eyes is not a different vantage of an image; it is a lack of a vantage at all. Second, like other reversible images whose shifting interpretation can be enhanced by movement of the eyes, the muscular adaptations required for seeing the alternate image of a Magic Eye stereogram is a necessary but not sufficient condition for its reinterpretation. This is illustrated by the fact that many people, despite all efforts of eye crossing, are incapable of entering them and that those who do have the good fortune to be of being able to experience them typically must engage in sustained cognitive effort to unpack the image once they begin to get into them. The central point of the Magic Eye example is that it illustrates how changing vantages on what one is looking at can profoundly influence what one believes to be true about it. The fact that this changing vantage may require a little eye crossing does not, in my view, lessen this observation.



Figure 5: The young girl-old woman illusion (otherwise known as “My Wife and My Mother-in-Law”) is a reversible image in which the viewer may either observe a young girl with her head turned to the right or an old woman with a large nose and protruding chin, depending on one’s perspective. Wright, E. (1992) The original of E. G. Boring’s Young Girl/Mother-in-Law drawing and its relation to the pattern of a joke. *Perception*, 21, 273–275.

The inception of psychology was marked by a concern with the inner experience of the individual (Schultz & Schultz 1992). Introspection was the tool of choice, and research entailed asking participants to scrutinize the components of their experiences. In short, psychology began with a fixed first-person perspective. In fact, it was during this time that psychology created some of its most robust laws of psychophysics demonstrating strikingly rigorous relationships between changes in various perceptual estimates (e.g., perceived brightness, weight, volume) and changes in the physical stimuli themselves (for a history, see Murray 1993). Then, concerns about the value of introspection arose, and researchers began to vacillate regarding the value of introspection relative to more “objective” third-person perspectives. Although

some researchers (notably the Gestalt psychologists and other researchers in the domain of human perception, e.g., Katz 1925/1989) continued to maintain a concern with inner experience, for a significant period of time the behaviorist reign caused a shift toward disregarding people’s first-person perspectives. Internal experience was a taboo topic. In short, psychology switched to a fixed third-person perspective. Then, with the rise of information processing and the cognitive era, the field again began to vacillate back and forth between considering people’s internal experiences and focusing on their behavior.

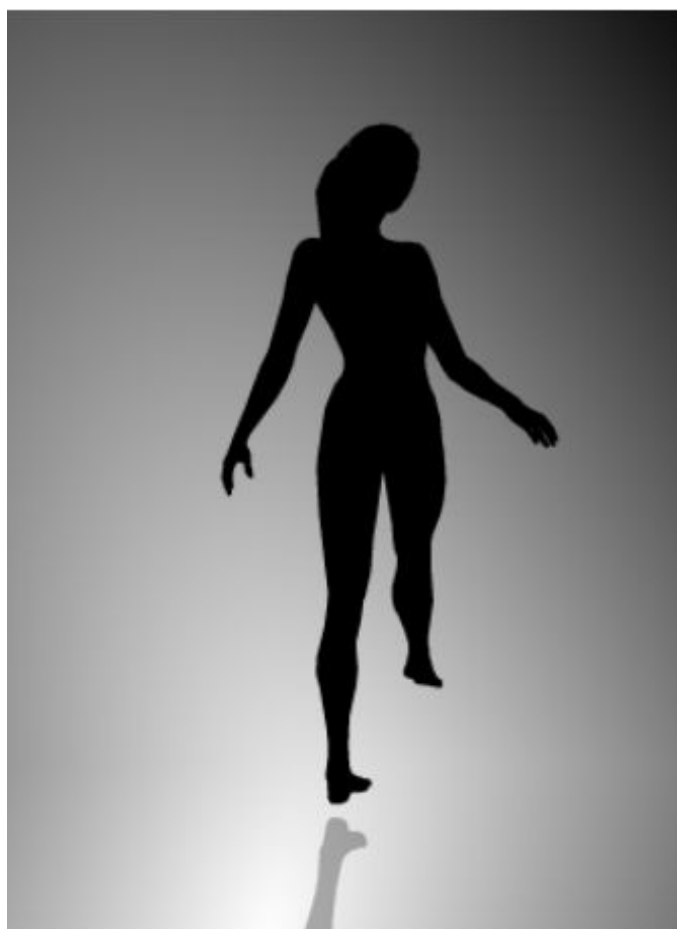


Figure 6: The spinning dancer illusion, or silhouette illusion, depicts a woman spinning in a circle. The direction of the dancer’s spinning (clockwise or counterclockwise) is dependent on the perspective taken by the observer. Kayahara, Nobuyuki (2003). Silhouette Illusion. *ProCrea*. Retrieved from <http://www.procreo.jp/lab0/lab013.html>

While psychology again finds itself in an age of flipping perspectives about first- versus

third-person accounts, much consternation still arises from this fact. Science in general (Wilber 1998) and psychology in particular (Wallace 2000) still find it challenging to fully integrate subjective experience into their accounts. Just as it is impossible to see a Necker cube simultaneously facing in its alternative directions, so too psychology has struggled to reconcile its vacillation between first- and third-person perspectives. On the one hand, ignoring the inner realm of experience seems to leave out much of “what it is like” to be human (Nagel 1974). On the other hand, researchers are rightly concerned about the validity and meaning of people’s first-person reports (Wilson 2003). With no alternative window into people’s minds, how can we know that their reports accurately correspond to their inner experience? After all, science necessarily relies on mutually agreed-upon observations. So how can we evaluate the first-person perspective that by its very nature eludes such consensus? The challenge is how to translate these first-person experiences into third-person data that can be scientifically investigated. The most straightforward answer of course is simply to ask people about their experience; their observable verbal statements thus become the third-person window onto their first-person experiences. But here we run up against the challenge that caused psychology to abandon the first-person perspective in the first place: How do we know if self-reports line up with first-person experiences without some independent measure of people’s internal states (Bayne this collection)?

Fortunately, self-reports are not the only third-person window into people’s inner experience. We can also examine other behaviors as well as measure physiological and brain activity in order to make reasoned inferences about what individuals are genuinely experiencing. In this manner, we can begin to discern when people are accurately characterizing their internal experience, and when they may be overlooking or distorting key aspects. The approach that I am advocating here is very much in keeping with Dennett’s notion of heterophenomenology (2003) that takes at its starting point the premise that people’s self-reports do not neces-

sarily reflect what they are actually experiencing but rather “*what the subject believes to be true about his or her conscious experience*” (Dennett 2003, p. 2). Although such an approach refrains from necessarily taking people’s first-person reports on face value, it does not abandon the prospect of making inferences about what people are actually experiencing.² Rather it posits that we must evaluate people’s self-reports in light of other third-person measures. As Dennett (1993) puts it:

My suggestion, then, is that if we were to find real goings-on in people’s brains that had enough of the ‘defining’ properties of the items that populate their heterophenomenological worlds, we could reasonably propose that we had discovered what they were really talking about—even if they initially resisted the identifications. And if we discovered that the real goings-on bore only a minor resemblance to the heterophenomenological items, we could reasonably declare that people were just mistaken in the beliefs they expressed, in spite of their sincerity. (p. 95)

As will be argued there are at least some situations in which external observers may have better knowledge of a person’s internal state than does the person in question. Moreover, there are some mental states (e.g., mind-wandering) for which the crucial bottleneck in people’s introspective awareness stems not from their capacity to classify the experience, but rather from the fact that people only intermittently take stock of what is going on in their own minds.

In the following section, I review some of the insights about first-person experience that can be gained when it is assessed from a third-

2 In the past (Schooler & Schreiber 2004) I characterized Dennett as dismissing the notion of underlying experience altogether, noting that he has written “Nobody is conscious... we are all zombies” (Dennett 1993, p. 406). Although I still find his views on this issue somewhat slippery, I now believe that he endorses the existence of genuine phenomenal experience that can be validated with third-person evidence. For example Dennett (2003) argues that evidence about briefly presented stimuli could help to inform subjects about their actual conscious experience observing “Subjects would learn for the first time that they were, or were not, conscious of these stimuli” (p. 9).

person perspective. By adopting a “trust but verify” approach to first-person reports, we not only gain a more objective understanding of subjective states, but also potentially glean a more astute perspective of our own experience.



Figure 7: “Magic Eye” stereograms can reveal a holographic three dimensional realm that is not initially perceptible at all. Consisting of abstract visual patterns constructed from an algorithm based on the principles of stereopsis, “Magic Eye” illusions require the viewer to blur their vision for a period of time, thereby revealing a three-dimensional imprint once perspective has shifted. Image provided by eyetricks.com. Additional examples can be found at <http://www.magiceye.com/3dfun/stwkdsp.shtml>.

3 Gaining a third-person perspective on people’s first-person experience

On some occasions we simply have experiences, but at other times we reflect on those experiences; that is, we intermittently take stock of our ongoing experience and re-represent it to ourselves. This distinction between having an experience (experiential consciousness) and explicitly re-representing it to ourselves (meta-awareness) is illustrated by the example of mind-wandering while reading (Schooler 2002). All of us have had the experience of reading along and suddenly realizing that, despite our best intentions, our eyes have been moving across the page but our minds have been entirely elsewhere. Indeed this has likely happened

to a goodly proportion of the readers whom have made it this far. The immediate question that this common experience raises is: why do we continue to simultaneously read and mind-wander even though we know that it is impossible to fully do both at the same time? The answer I suggest, and I’ll offer more evidence for this contention shortly, is that we routinely lose track of the contents of our own minds. People continue mind-wandering while reading because once they begin to mind-wander they often temporarily fail to notice (i.e., become meta-aware of) the fact that their minds are thinking about something unrelated to the text.

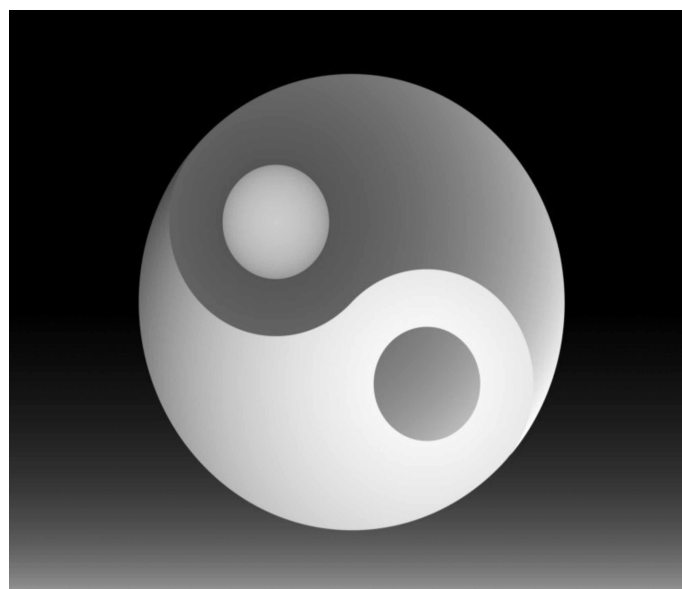


Figure 8: The three-dimensional image is a three dimensional yin-yang which the original Magic Eye image would not have revealed without a shift in perspective. The embedded three dimensional image has absolutely no resemblance to the two dimensional image from which it emerges. Image provided by eyetricks.com.

The notion that people routinely shift in perspective (from simply experiencing to attempting to re-represent their experience to themselves) provides the foundation for a framework for scientifically investigating first-person experience. Specifically, the distinction between experiential consciousness and meta-awareness raises the prospect of two types of dissociations between these vantages that are empirically tractable (Schooler 2002). *Temporal dissociations of meta-awareness* involve situ-

ations in which individuals engage in an experience without explicitly realizing that they are doing so. The example of temporarily failing to notice that one is mind-wandering is an example of a temporal dissociation. *Translation dissociations of meta-awareness* occur when one distorts or otherwise mischaracterizes their experience to themselves. Shouting “I am not angry” at the top of one’s lungs is an example of this latter dissociation. In the following discussion I briefly outline the empirical approach for exploring these two types of dissociations.

3.1 Temporal dissociations of meta-awareness

Although failing to notice that one is mind-wandering is a particularly apt example of a temporal dissociation of meta-awareness, there are numerous other examples of experiences that can temporarily go without being explicitly noticed, including unnoticed emotions (Lambie & Marcel 2002; Schooler & Mauss 2010) suppressed thoughts (Baird et al. 2013), and various mindless behaviors (Schooler et al. in press). Temporal dissociations of meta-awareness readily lend themselves to empirical investigation. Two approaches have proven effective in delineating situations in which people temporarily fail to notice a particular mental state: self-catching versus probe-catching and retrospective measures (Schooler et al. 2011).

The self-catch/probe-catch methodology pits two common self-report techniques against one another. Participants are asked to indicate every time they notice a particular mental state (e.g., mind-wandering). If an individual reports that they have just noticed themselves engaging in that mental state, then this is by definition a demonstration that the mental state has reached meta-awareness. Thus, self-caught episodes provide a straightforward measure of mental states of which individuals have become meta-aware. However, within this methodology, participants also periodically receive experience-sampling probes (Hurlburt & Heavey 2001) in which they are asked whether, at that particular time, they had been engaging in that mental state. If people are caught engaging in the state

before they notice it themselves (via self-catching), this provides a metric of episodes of that state that have eluded meta-awareness. As will be detailed later, this approach has proven effective in documenting temporal dissociations of a variety of different mental states including both mind-wandering (Schooler et al. 2004; Sayette et al. 2009; Sayette et al. 2010) and unwanted thoughts (Baird et al. 2013).

A second approach for identifying temporal dissociations of meta-awareness is to rely exclusively on experience sampling probes (i.e., probe-catching) but to additionally query people when they are caught in a particular state (e.g., mind-wandering) regarding whether or not they had been previously aware of that fact. Again, as will be seen, this strategy routinely reveals that people can be caught engaging in mental activities that they were previously experiencing but were not explicitly aware of. Intriguingly, the findings with this measure of temporal dissociation align with those revealed by the self-caught/probe-caught methodology to reveal consistent systematic differences between mental states associated with meta-awareness and those that lack it.

3.2 Translation dissociations of meta-awareness

Translation dissociations correspond to situations in which, while in the process of re-representation, one omits, distorts, or otherwise misrepresents one’s mental state to oneself and/or others. The basic strategy for assessing translation dissociations is to examine the correspondence between individuals’ self-reports of their mental states and indirect measures that might reasonably be expected to correspond to that state (Schooler & Schreiber 2004). If the correspondence is high, there is good reason to think that individuals are accurately reporting their internal state. If the correspondence is low, one needs to at least be suspicious that people are mischaracterizing their mental state.

Emotions are likely to be a particularly common source of translation dissociations. For example, when individuals report experiencing anxiety, a host of physiological measures (includ-

ing heart rate and galvanic skin response) typically become elevated (Marks 1987). Such correspondence gives us confidence that people are accurately characterizing their internal state; in other words, there is no translation dissociation. However, there is a class of individuals, referred to as repressors, who show the standard physiological changes when put in situations that would cause most people to experience anxiety, but who fail to report any change in anxiety (Asendorpf & Scherer 1983). In these cases, it seems reasonable to speculate that the repressors are misrepresenting their internal experience to themselves; they are experiencing anxiety but not acknowledging it (Lambie & Marcel 2002; Schooler et al. in press). As another example, consider that when males experience sexual arousal they typically show changes in their penile tumescence (a technical way of saying they become erect). Intriguingly, men who reported disgust for homosexual activity were shown to actually exhibit greater increases in penile tumescence when witnessing males engaging in sex, than men who did not report aversive feelings toward homosexuality (Adams et al. 1996). One reasonable account of these findings is that these so-called homophobics experience a translation dissociation, such that they are unable to acknowledge the arousal that they feel towards men, and instead misattribute the experience to a feeling of disdain.

A final example of translation dissociations involves situations in which individuals analyze why they feel the way they do about an affective experience. For example, in one study (Wilson et al. 1993), participants viewed various art posters and then both rated the posters and selected one to take home with them. Prior to engaging in this assessment, some participants were further asked to analyze why they felt the way they did about the posters, whereas others were not. When contacted several weeks later, people who had attempted to reflect on the basis of their preferences were less satisfied with their choice and were less likely to have hung the poster on their wall than those who had not analyzed their reasons. The disruptive effects of analyzing reasons, which have been conceptually replicated in a variety of contexts (Wilson & Schooler 1991), suggest that sometimes self-

reflection may be a source of translation dissociations. That is, in the process of trying to understand why people feel the way they do, they may construct a faulty meta-conscious representation and thereby lose touch with their feelings.

3.3 Investigating temporal and translation dissociations of meta-awareness in the context of mind-wandering

In recent years, a growing body of research has addressed the nature of mind-wandering as it pertains to the occurrence of temporal and translation dissociations of meta-awareness. This research suggests that mind-wandering is highly susceptible to temporal dissociations of meta-awareness; that is, individuals routinely fail to notice that their minds are wandering despite the considerable disruption to performance that such unnoticed lapses often incur. This claim is supported by various strands of evidence revealing the frequency with which participants are routinely “caught” mind-wandering before they notice it themselves. In contrast, mind-wandering appears to be relatively resistant to translation dissociations of meta-awareness. Although individuals regularly fail to notice when their minds are wandering, when meta-awareness is directed toward the current state of thought, they are generally quite accurate in characterizing whether or not their minds were on-task. This latter claim is supported by numerous demonstrations of systematic differences in performance and neurocognitive activity as a function of individuals’ self-classifications of their mental state as on-task versus mind-wandering.

3.3.1 On the veracity of self-reports of mind-wandering: How susceptible is mind-wandering to translation dissociations?

A fundamental challenge to the investigation of mind-wandering is its necessary reliance on self-report. Mind-wandering is, by its very nature, defined in terms of internal mental states. Given psychology’s long suspicions about introspective evidence (Nisbett & Wilson 1977), this reliance

on self-reports likely contributed to why, until recently, consideration of this important topic was largely limited to a few stalwart researchers (Antrobus 1999; Klinger 1999; Singer 1988; Giambra 1995). However, accumulating evidence suggests that when individuals are directly queried regarding whether they are mind-wandering, their self-reports accurately reflect their internal mental state. Evidence for this claim is largely based on the logic of triangulation (Schooler & Schreiber 2004). Accordingly, if self-reports of mind-wandering consistently co-vary with behavior and neuro-cognitive activity in a manner that might reasonably be expected to be impacted by mind-wandering, then we can have increased confidence that such introspective evidence accurately reflects the underlying mental state. In the following review, I detail at some length numerous findings in support of this relationship from a host of paradigms in which potential behavioral or physiological proxies of mind-wandering are related to individuals' responses to randomly timed queries regarding whether they were just mind-wandering. This review provides a review of the extensive literature on mind-wandering and evidence for the general contentions that: 1) the concordance between behavioral and physiological measures and self-report data indicate that people's self-reports of mind-wandering correspond to actual instances of this mental state; and 2) while people are routinely able to recognize mind-wandering after the fact, they often fail to notice it while it is occurring. Readers willing to take my word on these two points may want to scan or skip this section and jump ahead to its *Summary* (on page 16) or to the *Implications of this approach for the more general enterprise of the science of first-person perspective* (on page 18) if the general topic of mind-wandering is not of primary interest.

3.3.1.1 Behavioral measures

Reading comprehension

Although long overlooked as a source of reading comprehension failure, Schooler et al. (2004) found a strong correlation between the frequency of mind-wandering reports in response

to experience sampling probes and comprehension accuracy. Subsequent work demonstrated that mind-wandering specifically disrupts the development of a detailed situational model (Smallwood et al. 2008).

Another way in which the absence of reading comprehension following mind-wandering has been documented is through the examination of people's capacity to detect when the text becomes gibberish. In one study (Zedelius et al. 2014) participants were asked to read simple children's texts and report every time they noticed that the sentences no longer made any sense (some of the sentences were constructed so that the nouns of the sentences were rearranged in a nonsensical manner, e.g., "This sense makes no sentence"). The results revealed that participants sometimes continued reading for a number of sentences before noticing that the text had become gibberish. Moreover, participants who received thought probes after several sentences of gibberish were more than twice as likely to report mind-wandering without meta-awareness, relative to those who were probed at random times.

Eye-movements

If individuals' self-reported mind-wandering episodes during reading correspond to genuine mental lapses, then we might also reasonably expect to see differences between the patterns of gaze durations following periods in which individuals report reading attentively versus mind-wandering. These predictions were confirmed in an experiment in which subjects read the entirety of Jane Austen's *Sense and Sensibility* while their eye movements were recorded (Reichle et al. 2010). Relative to eye movements obtained during intervals of normal reading, the fixations measured during intervals that preceded reports of mindless reading were both longer in duration and less modulated by variables that are known to influence fixation durations (e.g., word frequency, Rayner 1998). These results suggest that the fairly tight coupling between the mind and eye during normal reading (Reichle 2006) becomes disengaged during self-reported mind-wandering.

Sustained Attention to Response Task (SART)

Another paradigm that has proven effective in documenting the validity of mind-wandering reports is the SART task. The SART is a simple go/no-go task in which participants are asked to refrain from responding to an infrequent no-go target (Manly et al. 1999; Robertson et al. 1997). Studies have documented that the brief lapses associated with this task share important features associated with reports of off-task thought. For example, individual difference measures such as cognitive failures (Smallwood et al. 2004), depression (Carriere et al. 2008; Farrin et al. 2003; Smallwood et al. 2007), and poor executive control (McVay & Kane 2009) have been associated both with greater mind-wandering reports and more errors on the SART. Similarly, both off-task reports and errors in this task share similar information processing features in terms of measures such as reaction time (RT) and evoked response potentials (ERPs; Smallwood et al. 2008, 2004, 2007).

3.3.1.2 Neurocognitive measures

Evoked Response Potential

When the brain faces situations in which it toggles between alternative perspectives, it routinely temporarily inhibits one perspective in favor of the other. This dampening of the non-dominant perspective is shown in reversible figures, where brain activation of one interpretation is inhibited while the other is consciously experienced (Tong et al. 2006). This same process of dampening the nondominant vantage also appears to operate when people favor their internal train of thought over external events. Accordingly, reports of mind-wandering should be associated with a dampening of attention to external stimuli. Indirect support for this “decoupling hypothesis” comes from studies demonstrating that participants are more prone to errors during periods associated with self-reported attentional drifts (e.g., Carriere et al. 2008; Smallwood et al. 2004; Weissman et al. 2006) and that they are less likely to recollect

external events during these periods (Smallwood et al. 2003, 2007, 2004).

More direct support for a relationship between self-reports of mind-wandering and dampened external processing comes from several ERP studies. In one study (Smallwood et al. 2008), participants intermittently received experience sampling probes while performing a simple target discrimination task. Analysis of the ERP responses to the targets revealed that the amplitude of the P3 ERP component elicited by the targets was significantly reduced for targets associated with “off-task” relative to “on-task” reports. Given that the P3 component reflects the degree to which external events are cognitively analyzed (e.g., Donchin & Coles 1988), these initial data support the proposal that mind-wandering reports are associated with an attenuation in stimulus processing at relatively late, post-perceptual processing stages.

A more recent ERP study examined whether mind-wandering might also attenuate sensory-level cortical processing (Kam et al. 2011). Participants again performed a simple discrimination task (at fixation) while being prompted at random intervals to report on their attentional state, but this time we also included task irrelevant probes in the visual periphery. The results revealed that the initial sensory-evoked response to probes was significantly attenuated prior to reports of “off-task” attentional states, as measured via the visual P1 ERP component. A second experiment that included irrelevant auditory probes similarly revealed that sensory-level auditory processing in the cortex is also dampened during self-reported “off-task” states, as measured via the auditory N1 ERP component. Another recent study from our lab (Baird et al. 2014) replicated the finding that mind-wandering reduced the P1 ERP, and further revealed that mind-wandering was associated with decreased phase-locking of electroencephalograph (EEG) neural oscillatory activity to sensory stimuli, suggesting that mind-wandering disrupts the temporal fidelity with which the brain responds to a stimulus.

Taken together, the collective ERP and EEG evidence demonstrates that self-reports of mind-wandering correspond to attenuated sens-

ory processing and cognitive appraisals of external stimuli. This finding further confirms the validity of self-reports of mind-wandering and suggests that a central feature of the mind-wandering state is an attenuation of the processing of external stimuli.

Functional magnetic resonance imaging (fMRI)

One of the challenges facing the burgeoning discipline of cognitive neuroscience is making sense of the observation that several brain areas, including the posterior parietal cortex and the precuneus, the medial prefrontal cortex, and the medial temporal lobe (which are collectively known as the default mode network (DMN), Raichle et al. 2001), all exhibit high levels of activity when participants have no external task to perform. One candidate process that the DMN could serve is the generation of the stimulus-independent thoughts that occur during the mind-wandering state, a hypothesis that is supported by a growing body of evidence. For example, McGuire et al. (1996) used the technique of retrospective thought sampling to demonstrate that reports of mind-wandering were associated with activity in the medial prefrontal cortex. More recently, several studies have documented that situations associated with greater mind-wandering reports (as assessed outside of the scanner) also lead to greater activity in many of the key elements of the DMN (Mason et al. 2007; McKiernan et al. 2006).

While activity in the DMN is correlated with high probability of retrospective reported mind-wandering, it was originally unclear whether particular episodes of self-reported mind-wandering are linked to recruitment of the DMN. To assess whether this was the case, we conducted a study in which experience sampling was combined with fMRI to assess the neural activity that occurred during particular episodes of mind-wandering (Christoff et al. 2009). This study revealed that, in addition to the activation of several core structures in the DMN, areas normally observed in controlled processing (including the dorsolateral prefrontal cortex and the dorsal

anterior cingulate) were also engaged during self-reported off-task thought. This pattern of brain activation suggests that executive and default network resources are jointly recruited during episodes of mind-wandering. One possible account explaining this joint activation is that executive network resources play a role in transforming the self-referential content supported by the DMN into the internal train of thought that we experience when the mind wanders. Further support for this hypothesis is provided by evidence that the ability to engage in autobiographical planning (such as “how do I get out of debt?”) requires cooperation between the DMN and a system involving attentional control (Spreng et al. 2010).

Christoff et al. (2009) also compared the pattern of activations associated with introspective reports of mind-wandering, on the one hand and the pattern of activations associated with behavioral errors, on the other hand. Although a variety of factors are known to contribute to behavioral errors during the SART, mind-wandering is believed to be one important source of such errors. Consistent with this view, SART errors (Figure 9) and the introspective reports of mind-wandering (Figure 10) were associated with similar patterns of brain recruitment, providing further validation for the use of introspective experience sampling reports for the study of mind-wandering.

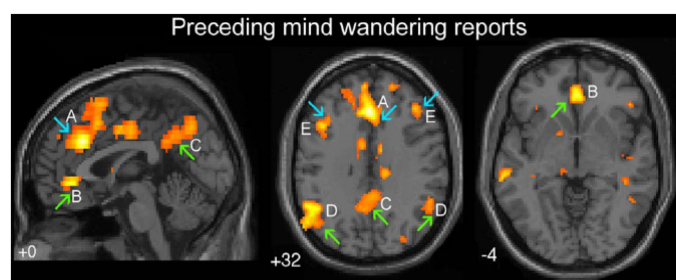


Figure 9: Activations preceding reports of mind wandering (off-task versus on-task). Upward green arrows: default network regions; downward blue arrows: executive network regions. Regions of activation included (A) Dorsal ACC (BA32); (B) Ventral ACC (BA 24/32); (C) Precuneus (BA7); (D) Left temporoparietal junction (BA 39); (E) Bilateral DLPFC (BA 9). Height threshold $P < 0.005$, extent threshold $k > 5$ voxels (from Christoff et al. 2009).

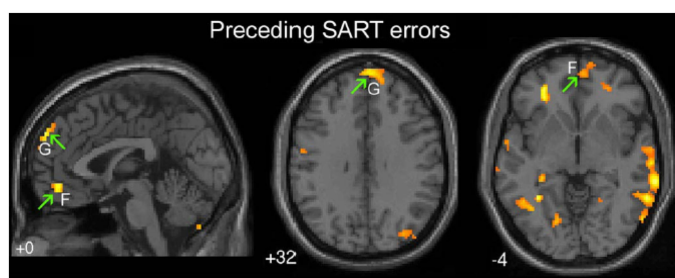


Figure 10: Activations preceding SART errors (interval prior to incorrect versus correct targets). Upward green arrows: default network regions; downward blue arrows: executive network regions. Regions of activation included: (F) Ventromedial PFC (BA10/11); (G) Dorsomedial PFC (BA9). Height threshold $P < 0.005$, extent threshold $k > 5$ voxels (from Christoff et al. 2009).

3.3.2 The intermittent meta-awareness of mind-wandering: How susceptible is mind-wandering to temporal dissociations?

Although when queried individuals are quite reliable in their capacity to self-report whether or not they were mind-wandering, a variety of strands of evidence suggest that people routinely fail to spontaneously notice when mind-wandering takes place. Two paradigms, reviewed earlier, have documented the intermittent meta-awareness of mind-wandering.

3.3.2.1 Self-caught/probe-caught methodology

One approach for documenting mind-wandering in the absence of meta-awareness is combining self-catching and experience sampling measures into a single paradigm. Recall that the self-catching measure asks participants to press a response key every time they notice for themselves that they have been mind-wandering. This measure provides a straightforward assessment of the mind-wandering episodes that have reached meta-awareness. The experience sampling measure, on the other hand, randomly probes people regarding whether they were at that particular moment mind-wandering. When used in conjunction with the self-caught measure, experience sampling can catch people mind-wandering before they notice it themselves.

A number of studies have effectively used the self-caught/probe-caught methodology to illuminate the relationship between mind-wandering and meta-awareness. This approach was initially used to examine mind-wandering while reading (Schooler et al. 2004) and revealed that whereas participants regularly caught themselves mind-wandering, they nevertheless were often caught mind-wandering by the probes. Strikingly, and in support of a fundamental difference between mind-wandering episodes that are accompanied by meta-awareness and those that are not, there was a strong correlation between probe-caught mind-wandering and comprehension performance but no such relationship between self-caught mind-wandering and comprehension.

Additional studies have examined the impact of two mind-altering experiences hypothesized to undermine individuals' meta-awareness: alcohol intoxication and cigarette craving. In one study (Sayette et al. 2009), social drinkers consumed a moderate dose of alcohol or a placebo beverage and then performed a reading task (implementing a self-caught/probe-caught mind-wandering assessment methodology). Compared with those who drank the placebo, participants who drank alcohol were more likely to report that they were “zoning out” when probed. After accounting for this increase in mind-wandering, alcohol also lowered the probability of catching oneself zoning out (i.e., self-catching). These data suggest that alcohol increases mind-wandering while simultaneously reducing the likelihood of noticing one's mind-wandering.

In another study (Sayette et al. 2010), smokers, who were either nicotine-deprived (crave condition) or non-deprived (low-crave condition), performed the same mind-wandering task used in Sayette et al. (2009). Smokers in the cigarette-crave condition were significantly more likely than the low-craving smokers to acknowledge that their mind was wandering when they were probed. When this more-than-threelfold increase in zoning out was accounted for, craving also lowered the probability of catching oneself mind-wandering. Similar to the alcohol consumption findings, it appears that ci-

garette craving simultaneously increases mental lapses while reducing the metacognitive capacity to notice them.

3.3.2.2 Retrospective classification of mind-wandering episodes

A second methodology that has been used to examine fluctuations in meta-awareness of mind-wandering entails combining the experiential sampling methodology with a judgment of participants' immediately prior state of awareness. Recall that, in the experience sampling procedure, participants are intermittently queried regarding whether or not they were mind-wandering; in this combined approach, if they report mind-wandering to the probe, then they are also asked to indicate if they were aware that they were mind-wandering. In response to such queries, participants routinely indicate that they had been unaware of their mind-wandering up until the time of the probe. Moreover, when participants classify mind-wandering episodes as unaware, their performance and neurocognitive activity systematically differ from when they report having realized they were mind-wandering.

Consistent with findings using the self-caught/probe-caught methodology, retrospective classifications of unaware mind-wandering episodes (termed zoning out) and aware episodes (termed tuning out), indicate that the former are more associated with comprehension failures than the latter (Smallwood et al. 2008). By contrast, reports of zoning out seem to be most closely linked to failures in response inhibition (Smallwood et al. 2008, 2007) and to poor mental models during reading (Smallwood et al. 2008). Together these results suggest that while maintaining streams of stimulus-independent thought interfere with the integrity of external attention, the absence of awareness of mind-wandering is especially damaging to task performance.

Neurocognitive measures also reveal differences in the degree of activation between mind-wandering episodes that have been classified as aware versus unaware. In the combined experience sampling/fMRI study conducted by Christoff et al. (2009), mind-wandering with awareness activated similar brain regions to those observed

during mind-wandering without awareness. These brain regions, however, were more strongly activated when mind-wandering occurred without awareness (see Figure 11). The anterior prefrontal cortex (BA10) was one of the brain regions significantly more strongly recruited during unaware episodes of mind-wandering. Notably, anterior prefrontal cortex (PFC) recruitment has been directly linked to engagement of cognitive meta-awareness (McCaig et al. 2011). The observation that this same brain region became specifically more recruited during unaware episodes of mind-wandering may seem surprising at first. However, the anterior PFC may be involved in mind-wandering through its role in the maintenance of thought. As discussed further below, its recruitment during mind-wandering in the absence of awareness may make it more difficult for meta-awareness to be implemented.

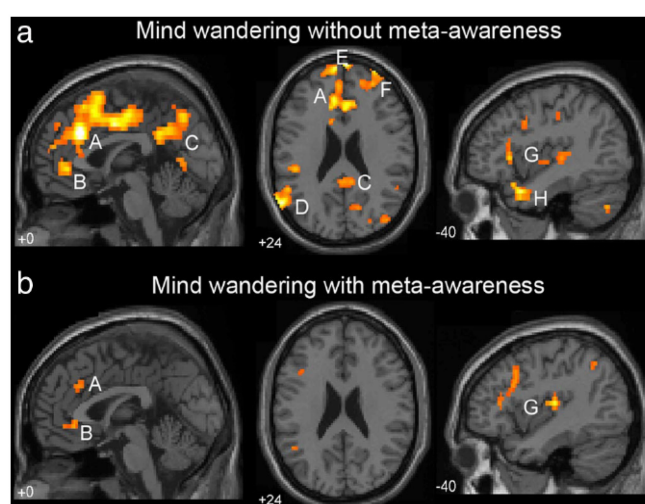


Figure 11: Mind-wandering in the (a) absence and (b) presence of meta-awareness. (a) Regions of activation associated with mind-wandering in the absence of awareness (off-task unaware versus on-task): (A) Dorsal ACC (BA32); (B) Ventral ACC (BA32); (C) Precuneus (BA7); (D) Posterior Temporoparietal Cortex (BA39); (E) Dorsal Rostromedial Prefrontal Cortex (BA10); (F) Right Rostrolateral Prefrontal Cortex (BA10); (G) Posterior & Anterior Insula; (H) Bilateral Temporopolar Cortex; (b) Similar regions were activated during mind-wandering with awareness (off-task aware versus on-task comparison) but to a lesser degree, including: (A) Dorsal ACC (BA32); (B) Ventral ACC (BA24/32); (G) Posterior & Anterior Insula. Height threshold $P < 0.005$, extent threshold $k > 5$ voxels (from Christoff et al. 2009).

3.3.3 Summary

In sum, the investigation of mind-wandering from the vantage of the distinction between having an experience (experiential consciousness) and explicitly realizing that one is having an experience (meta-awareness) has provided a fertile ground for developing a third-person understanding of first-person experience. This research has begun to chart the stream of consciousness, demonstrating that individuals regularly vacillate between the outer realm of perception and the inner realm of thoughts and feelings. This fluctuation routinely evades explicit meta-awareness, enabling people's minds to move on to a new topic without explicitly realizing this fact. Nevertheless, when directly queried, people are remarkably capable of introspecting and noticing whether or not they were mind-wandering. The fluctuation of perspectives on the mind that this approach affords raises numerous questions. Here, I address three: 1) If people are so competent at recognizing that they are mind-wandering when queried, then why do they find it so difficult to notice this fact on their own? 2) Are there ways of enhancing the capacity to catch one's mind in flight? 3) What are the implications of this approach for the more general enterprise of the science of first-person perspective? I consider these questions in turn.

3.3.3.1 Why is mind-wandering so easy to report but so difficult to catch?

The observation that meta-awareness is so effective at discerning mind-wandering when queried about it, yet so poor at catching it on its own, raises the natural question of why this discrepancy exists. Two potentially interrelated explanations may contribute to this striking discrepancy.

Like mind-wandering, meta-awareness appears to be associated with rhythms of attentional flux (Schooler et al. 2011). Sometimes we are explicitly aware of our mental states, and other times we are not. Such vacillations in meta-awareness could readily contribute to individuals' frequent tendency to overlook episodes

of mind-wandering, as this mental state may only be notable when the explicit spotlight of attention is metaphorically turned on itself. Indeed the tendency to only notice mind-wandering after the fact may similarly apply to other mental states that routinely curtail the occurrence of meta-awareness. Like mind-wandering, other subjective states such as sleep, anesthesia, dreaming, and flow states are typically not noticed while they are occurring, but are readily acknowledged after the fact. Sleep (in the absence of dreaming) and anesthesia are typically lacking conscious experience entirely and so clearly are not candidates for meta-awareness. The mental states associated with gradually drifting off to sleep and dreaming do have phenomenal content but typically lack meta-awareness. This is why people routinely don't notice that they are falling asleep (a grave danger for driving) or dreaming (except in the case of lucid dreaming, LaBerge 1980). Another example is that of flow states (Csikszentmihalyi 1988), during which people engage in highly demanding tasks at close to their optimum level of performance. In such cases, people lack the additional resources to take stock of their experience, which may be why meta-awareness of a flow state often leads to its sadly premature termination. Nevertheless, as in the other cases, after a flow state has ended, individuals are quite able to acknowledge its occurrence. In all of these cases, the common denominator may be that these various states (for one reason or another) curtail the occurrence of meta-awareness, and thus are only noticed after the fact once the opportunity for meta-awareness reoccurs.

One reason why mind-wandering may undermine meta-awareness may stem from its reliance on the very same brain regions that may be necessary for noticing its occurrence. A striking aspect of the brain regions associated with mind-wandering is that they involve many of the systems that might be expected to contribute to the monitoring of the state. For example, elements of the medial prefrontal cortex are recruited both during mind-wandering and in tasks that require theory of mind (Gallagher & Frith 2003). As mental state attribution involves the application of meta-cognitive pro-

cesses to information of a stimulus-independent nature (e.g., inferences about the mental state of another individual), the engagement of these brain regions during mind-wandering could prohibit their utility in the service of catching the wandering mind. Similarly, in the combined fMRI/experience sampling study conducted by [Christoff et al. \(2009\)](#), periods of mind-wandering engaged regions such as the dorsal ACC, involved in error-detection and conflict monitoring, and the anterior PFC, involved in cognitive meta-awareness. If mind-wandering engages both meta-cognition and error-detection systems in the service of generating a coherent stream of stimulus-independent thought, the fact that these systems are already engaged may make them less capable of detecting a mind-wandering episode. The observation that mind-wandering and meta-cognitive processes both engage the same systems does not necessarily establish a causal relationship between these two. Nevertheless, it remains an intriguing speculation that our persistent failure to catch ourselves mind-wandering may occur because mind-wandering hijacks the precise meta-cognitive brain regions that are necessary for noticing it. Future research might profitably explore this hypothesis by examining whether mind-wandering episodes that are experimentally induced to emphasize meta-cognitive reflection are particularly likely to evade detection.

3.3.3.2 Are there ways of enhancing people's awareness of their mind-wandering?

One of the clear findings of research on mind-wandering is that it can be extremely disruptive to performance. Reading ([Smallwood et al. 2008](#)), working memory ([McVay & Kane 2009](#)), vigilance ([Cheyne et al. 2009](#)), and general intellectual functioning ([Mrazek et al. 2012](#)) can be seriously disrupted by mind-wandering, especially when it occurs without awareness ([Smallwood et al. 2008](#)). This raises the natural question of whether enhancing people's meta-awareness of their minds can help to curtail the disruptive consequences of mind-wandering.

Of course, just because episodes of mind-wandering routinely end with a moment of meta-awareness ("shoot, I drifted off again") does not mean that the meta-awareness necessarily was responsible for its ending ([Schooler et al. 2011](#)). Meta-awareness could be a consequence rather than the source of the termination of a mind-wandering episode. According to this view, the intuition that meta-awareness terminates mind-wandering episodes is another example of an over-reach of the attribution of deliberate intention ([Metzinger 2013](#)). While this remains a viable possibility, it is also the case that mindfulness techniques aimed at enhancing awareness of one's internal states can curtail the negative effects of mind-wandering.

In one study ([Mrazek et al. 2013](#)), participants were randomly assigned to one of two interventions that they were told were expected to enhance their performance: two weeks of training either in mindfulness meditation, or in good nutrition practices. Both interventions involved similar time commitments, expectations, and homework (either daily meditation or a nutrition journal). Before and after the intervention, participants were given both reading comprehension and working memory tasks, and their mind-wandering during each was assessed. Compared to the nutrition control, the mindfulness intervention significantly reduced mind-wandering, improved performance on both tasks, and these benefits were mediated by the reduction in mind-wandering for those who were high in mind-wandering to begin with. These findings dovetail with other recent studies indicating that the general tendency for mindfulness (being present in the moment) is negatively correlated with mind-wandering ([Mrazek et al. 2012](#)), and that even a simple mindfulness exercise conducted with non-meditators (focusing on one's breath for eight minutes) can temporarily reduce mind-wandering ([Mrazek et al. 2012](#)).

Although research on the impact of mindfulness training in dampening mind-wandering is consistent with the notion that part of its efficacy is due to enhancing meta-awareness, there is one finding that does not completely square with this account. Specifically, [Mrazek et al. \(2012\)](#) found that mindfulness training re-

duced people's tendency to spontaneously notice mind-wandering episodes. However, this reduction in self-caught mind-wandering could have occurred because the mindfulness practice enhanced people's awareness of the focus of their attention, thereby preventing them from initiating mind-wandering episodes in the first place. Consistent with this speculation, another recent study (Baird et al. in press) demonstrated that a similar mindfulness program can enhance at least one meta-cognitive skill, namely, the ability to assess the accuracy of memory recognition judgments. Although more research is clearly needed, it remains quite plausible that one mechanism by which mindfulness training reduces mind-wandering is by increasing people's meta-awareness of when their minds are beginning to wander.

3.3.3.3 What are the implications of this approach for the more general enterprise of the science of first-person perspective?

The program of research outlined above demonstrates the insights into first person experience that can be gleaned by assessing it from a third-person perspective. In many respects, the approach described here exemplifies the program of heterophenomenology that Dennett advocates. We are systematically assessing people's reports about their conscious experiences while explicitly acknowledging that those reports correspond to people's beliefs about their experience (i.e., their meta-awareness) and not necessarily their actual experience. However, by using various reasonable markers of people's internal states we have been able to examine the conditions under which people's reports are more or less likely to be aligned with their experience. In this regard, we find that when people are explicitly asked whether they were just mind-wandering, their self-reports align with a host of behavioral and physiological measures that should co-vary with mind-wandering. These findings suggest that people are quite accurate in retrospectively assessing whether or not they were just mind-wandering. In other words, by triangulating between people's retrospective self-re-

ports of mind-wandering (following experience sampling cues) and both behavioral and physiological measures, we have identified situations in which all evidence suggests that people's opinions about the content of their private experience is generally quite accurate.

At the same time, by introducing the self-caught procedure in combination with retrospective assessments of people's awareness of prior states of mind-wandering, we have also documented critical lacunae in people's knowledge of their mental states. Specifically we find that people routinely fail to spontaneously notice when their minds have wandered. When tasked with reporting mind-wandering whenever they become aware of it, people routinely demonstrate behavior indicative of mind-wandering while failing to report it. If they are probed during periods in which these measures suggest they are mind-wandering, they routinely indicate that they now realize that they were mind-wandering, but they had not noticed this state until the time of the probe. We are thus also able to identify situations in which all evidence suggests people are routinely lacking in their current knowledge of their ongoing mental state.

By triangulating between people's first-person reports and multiple other third-person measures we have begun to reveal the relationship between people's beliefs about their experience and empirical indices of their underlying mental states (for related approaches, see Hurlburt & Heavey 2001; Jack & Roepstorff 2002; Lambie & Marcel 2002; Lutz & Thompson 2003). Moreover, the theory of the intermittent and imperfect nature of meta-awareness as a representation of experience (Schooler & Schreiber 2004; Schooler 2002; Schooler et al. 2015) provides a scaffold for conceptualizing the situations in which beliefs and underlying experience converge and diverge. Of course, one could always counter that we cannot be sure that the variety of behavioral and physiological measures that correlate with self-reported mental states such as mind-wandering are necessarily indicative of those states. Perhaps there is some third variable that is responsible for both mind-wandering and the host of measures that

we find to be correlated with people's self-reporting of it. But it seems a stretch to suggest that this entirely unknown third variable could account for why, when people say they were mind-wandering, their performance on primary tasks is impaired, their eye movements become less sensitive to what they are looking at, their physiological measures indicate a dampening of attention to external processes, and their brain activation corresponds to that which occurs when they are unoccupied. In short, a strong case can be made for the value of using empirical third-person science to inform not only our understanding of people's beliefs about their experience, but also to discern when those beliefs are likely to be accurate and when they may be inaccurate or incomplete.

It seems likely that those with strong allegiances to either an exclusively first- or third-person account of experience will balk at the notion that third-person empirical indices can be used to corroborate people's first-person accounts. Traditional phenomenologists (e.g., [Husserl 1963](#)) may contend that first-person experience is privileged and so, when discrepancies arise between it and third-person data, that the former should invariably be favored. Those with a behaviorist bent may argue that making claims about underlying subjective states remains a dead end because ultimately they can never truly be verified. Personally I find myself sympathetic to both of the vantages; however, I argue that the striking disparity of these views, both from each other and from the one promoted here, stems from the incongruence that naturally arises from shifting perspectives.

From the vantage of one perspective of a Necker cube, the alternative perspective makes little sense. When the spinning dancer is moving in one direction, it is hard to imagine how she could possibly shift directions. Those who have never entered the third dimension of a Magic Eye image could reasonably doubt that such a perspective could possibly exist. But once one realizes that there are distinctly different perspectives to be had on a situation, and that these alternative perspectives each offer their own valuable vantage, then that knowledge can be held even as one remains incapable of experi-

encing both at the same time. I believe this is the case with interpreting scientific third-person accounts of first-person experience. If one is capable of recognizing both the strengths and limitations of each perspective, then they can use each to inform the other. If, however, they solely look at a problem from one or the other perspective, then this may lead to a logically consistent view, but one that omits an important vantage. I turn now to a consideration of this larger issue: namely, conceptualizing a meta-perspective that can accommodate the vacillating manner in which first-person experience is both that which we know best and understand least.

4 Toward a meta-perspective for considering the metaphysics of first-versus third-person perspective

It is my contention that debates about how to reconcile the first- and third-person perspective on reality arise in part from the distinct vantages that different scholars take on the issue. The problem in a nutshell is that while the prevailing third-person perspective of science (material reductionism) does an admirable job of accounting for all aspects of reality that are revealed from its vantage, it robustly fails to accommodate several self-evident aspects of existence that are uniquely apparent from a first-person perspective. If one simply dismisses those aspects of the first-person perspective that are incongruent with the third-person perspective, (as most scientists and many philosophers do), then there is no problem. However, here I will argue that there exist self-evident observations derived from the first-person perspective that are as compelling as any objective fact. Such observations should not be simply dismissed as irrelevant or illusory but rather suggest the need of serious revision to current accounts of physical reality (for related arguments see [Chalmers 2002](#); [Nagel 2012](#)). In the following section, I first review the material reductionist account suggested by the prevailing third-person perspective view. I then consider several elements of existence revealed by a first-person perspective that seem to have no place

in this account, most notably subjective experience, the flow of time, and the distinctiveness of the present. Finally, I offer some speculative remarks about the nature of a meta-perspective that might be able to accommodate both vantages.

4.1 Ontological third-person perspective—Material reductionism

When reality is conceived of strictly from the vantage of a third-person perspective, it quite naturally leads to the premise of material reductionism, namely that everything including the arising of subjective experience can be accommodated on the basis of physical principles that do not themselves make any appeal to consciousness. This account is arguably the prevailing view among both scientists (e.g., [Crick 1994](#); [Bloom 2009](#); [Graziano 2013](#)) and philosophers (e.g., [Dennett 1993](#); [Churchland 1989](#); [Metzinger 2004](#)). Its strength comes from its remarkable record of success. Having abandoned the superstitions and spiritual whimsies of the past, hard-nosed science has an amazing track record for explaining everything it has been directed toward with purely physical constructs. Aspects of reality that were once thought to be beyond the ken of the third-person perspective of science, for example the notion of some sort of mystical force of life, *élan vital*, have been reduced to rigorous formalisms (e.g., DNA code). Admittedly, we do not currently have a full accounting of how it is that we experience a first-person perspective on reality, but given science's track record, it is presumed to be merely a matter of time before these experiences are explained with precisely the same type of accounts that have been used so successfully to explain so much so far ([Churchland 1989](#)). People may feel as if they have some type of privileged perspective, as if the view from within their own minds could never be reduced to and explained by the machinations of atoms, but this is just shortsightedness, perhaps fueled by some evolutionary advantage to view mind and matter as different ([Bloom 2009](#)).

There is much to be said for material reductionism, as it draws on the very assumptions

that have led to the remarkable progress of science. To appeal to the existence of some other distinct realm of reality beyond the objectively physical smacks of ghosts and fairy dust (e.g., [Jackson 1982](#)). To date, while the previous analysis has revealed the marked advances to our understanding that emerge when we consider people's first-person perspectives, no *explanation* in science has required abandoning an exclusive reliance on mutually verifiable third-person observations. In other words, although I will soon suggest cases that may challenge this tradition, to date there are no third-person accounts of physical phenomena that have been undermined solely because they conflict with first-person experience. Given the track record of third-person accounts, it may seem hard to justify why one scientific question (the arising of conscious experience) should challenge an ontological perspective that is not problematic for anything else.

4.2 Ontological first-person perspective—What material reductionism leaves out

Although material reductionism provides an outstanding vantage for accounting for the physical world, it comes up wanting when the mind is inspected from a first-person perspective. The essential challenge is that even if a materialistic explanation is able to account for how the mind functions, this does not explain how it is that there is a subjective experience associated with it, or why that experience is as it is. As [Jackson \(1982\)](#) puts it:

Tell me everything physical there is to tell about what is going on in a living brain, the kind of states, their functional role, their relation to what goes on at other times and in other brains, and so on and so forth, and be I as clever as can be in fitting it all together, you won't have told me about the hurtfulness of pains, the itchiness of itches, pangs of jealousy, or about the characteristic experience of tasting a lemon, smelling a rose, hearing a loud noise or seeing the sky. (p. 127)

Jackson introduces the canonical example of Mary the color scientist to illustrate this point. Imagine that Mary is a color scientist who has been brought up in a black and white room and has never experienced red; nevertheless, she knows all there is to know about the physical processes relevant to color vision. Jackson's point is that if she later experiences red firsthand, she will learn a new fact (the experience of red) that all of her physical knowledge was insufficient to provide. Complete physical knowledge about a subjective experience is insufficient to entirely know all there is to know about that experience. One has to actually have the first-person experience to fully understand it.

A second criticism of material reductionism involves its inability to explain the arising of conscious experience. It is quite straightforward to imagine how physical processes could account for the structure and function of the mind in much the same way that they can explain the structure of computer hardware and the functions of computer software. But such an account would not explain how subjectivity itself arises or what it is like from the vantage of the experiencer. Similarly, even if we were to create a computer that perfectly emulated a conscious being, we could not know whether it was genuinely conscious, and if it were, "what it is like to be" (Nagel 1974) a computer.

The inherent difficulty of conceptualizing how material objects enjoy subjective experience is further illustrated by a third criticism of material reductionism, namely that it is possible to conceive of a system that has all of the physical characteristics of a conscious being, but nevertheless lacks consciousness. Philosophical zombies (Chalmers 1995) are hypothetical human beings who have no internal experience but are otherwise identical to normal people in all other physical measurements and behaviors (including claiming that they are conscious). Although there is no way of demonstrating that such creatures could ever exist, there is also no way of demonstrating that they couldn't. Finding the neural correlates of consciousness helps not an iota, as even a zombie who reported consciousness in certain brain states would still not be actually enjoying a genuine experiential

state. If zombies that are physically indistinguishable from experiencing humans could in principle exist, then there is nothing inherent in what is known about physical systems that speaks to the arising of consciousness. This presents a major problem to the prevailing material reductionist view because it offers no way to distinguish between philosophical zombies and the non-zombies.

The essential problem of the exclusively third-person perspective of material reductionism is that it is forced to ignore all aspects of experience that cannot be reduced to a third-person perspective. A thought experiment may help to provide a further "intuition pump" (Dennett 2014) for illustrating just how special that extra something might be. Consider the following science fiction variant on the classic Faustian bargain (Goethe 1867). One day, to your amazement, a flying saucer lands in front of you and a member of a clearly more advanced species emerges and says that he/she (it's unclear) has been enjoying our debates about the mind-body problem, which his/her civilization has solved. If philosophical zombies are logically³ possible, you can be turned into one. He/she offers you all the gold you can imagine (they've also mastered alchemy) if you are willing to accept the risk of becoming a zombie. If a zombie is a logical possibility, you will be transformed into one. From everyone else's perspective (i.e., the third-person perspective), you will be exactly as you were before (just much richer). However, you will not actually have any experience at all; you will simply seem to others as if you do. Would you take the bargain? Hard-nosed material reductionists say they would (D. Dennett, personal communication, 7/15/2014; M. Graziano, 6/10/2014, personal communication), but many of the rest of us might not. What is the value of untold wealth, if there is no inner experience by which it can be enjoyed?

The *Zombie Faustian Bargain* serves as a useful intuition pump for illustrating the im-

³ Let's just assume, for the sake of argument, that the aliens had solved the tricky issue of moving from logical to nomological possibility, that is that if it is possible for a philosophical zombie to exist in any conceivable universe, that it would be possible for you to become one.

portance of the extra something that is left out of the third-person material reductionist perspective. Nevertheless, it is clearly a fanciful proposition and material reductionists might reasonably argue that there is not much to worry about if the only cost to adopting their view is not knowing how to respond to such an unlikely scenario. However, there are numerous other examples closer to home where the limits of a third-person accounting of consciousness become relevant. Issues surrounding the nature and existence of consciousness in other species, fetuses, and computers all revolve around inferences about first-person experiences that gravely exceed all known or conceived ways of reconciliation.

A less obvious domain for a clash between the current prevailing third-person perspective of science and first-person experiences arises in, of all places, physics. Although there has been some speculation, now largely disregarded by the mainstream, that consciousness could have something to do with the collapse of the wave function in quantum physics (Wigner & Margenau 1967), in general, consciousness is assumed to have little relevance to physics. However, there are two current assumptions in physics that seem to squarely contradict first-person experience. Specifically, physicists believe that the flow of time is an illusion and that there is nothing special about the present. Before considering why these claims are so problematic for the existence of subjective experience, let us first consider why physics makes this claim.

4.3 Why physicists dismiss the flow of time and the privileged present

In considering the nature of time, physicists often “spatialize” it. In other words, they attempt to place it on a similar footing to the traditional three dimensions of space (see Figure 10). Though differing from spatial dimensions in important respects (Einstein 2001), the notion of time as similar to a spatial dimension is a key feature of the prevailing Einstein/Minkowski interpretation of special relativity theory. Space and time are combined in this theory into one concept: space-time. The spatialization of time

allows the depiction of a “block universe” in which the traditional spatial dimensions are reduced (for purposes of visual illustration) to two dimensions from three, and time is added as a third dimension. Such a depiction can be thought of as a space-time “loaf of bread,” where each narrow cross-section of the loaf (“slice”) constitutes a moment in time of the entire universe. According to the block universe view (widely held by today’s physicists), all slices—past, present, and future—already exist. This arises from the relativity of simultaneity, which means that “now” is different for different observers. It is simply that each individual observer is privy to only one moment (slice) at a time. From the vantage of a block universe, the only thing that seems to actually move in time is consciousness itself (i.e., the observer). This means that from the vantage of the prevailing view of physics, the flow of time is not a part of objective reality but simply an artifact of subjective experience. As Stanford physicist Linde (2004) notes: “Thus we see that without introducing an observer, we have a dead universe that does not evolve in time” (p. 25). What is more, once we conceive of the temporal dimension as the equivalent of another spatial dimension, then there are not enough degrees of freedom for the observer to move in time; that is, movement requires a rate in time, but time in the block universe is already represented as a spatial dimension, and thus cannot also be used as the metric that establishes the rate of movement through time. As the physicist Paul Davies (2002) puts it:

Nothing other than a conscious observer registers the flow of time. A clock measures durations between events much as a measuring tape measures distance between places; it does not measure the ‘speed’ with which one moment succeeds another. Therefore it appears that the flow is subjective, not objective. (p. 36)

The upshot of this reasoning is that the flow of time is an illusion, an artifact of consciousness. Again, as Davies (2002) puts it: “From the fixed past to the tangible present to the undecided

future, it feels as though time flows inexorably on. But that is an illusion” (p. 32).

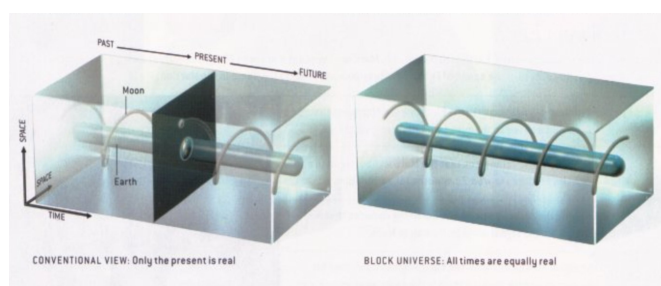


Figure 12: Although the conventional view derived from experience is that the present is real and moves through time, current views in physics say this is erroneous. According to the standard block universe view in physics, all moments—past, present, and future—are equally real. The flow of time and the privileged present are seen as illusions of consciousness (from [Davies 2002](#)).

The characterization of reality as a block universe, with the flow of time as an illusion of consciousness, also leads to the conclusion that the privileged present is an illusion. One of the most pronounced aspects of consciousness is its extension in time. Consciousness extends in time and thereby gains the “now” in which it resides. First-person observers may remember the past or imagine the future (as often happens during mind-wandering) but ultimately mental time travel always takes place in the present. The observer perpetually and exclusively resides in the present. In this sense, it seems intuitively self-evident that the “now” is privileged. But not so from the current vantage of the block universe in physics, where the present is treated exactly the same as the past and the future. As Einstein himself observed, “The past, present and future are only illusions, even if stubborn ones” (quoted in [Hoffmann & Dukas 1972](#), p. 258). Again, the problem is that the only thing that defines the present from the vantage of a block universe is that it is where the observer perceives itself to be at any particular moment in time. But from the vantage of a block universe, all moments of time exist simultaneously.

The notions that the flow of time and the privileged present are merely illusions of consciousness are less problematic from a third-person

perspective than the first-person perspective. If there is no ultimate reality to subjectivity, then there is no problem making claims that are directly in opposition to subjective experience. At a recent public lecture, I asked the noted physicist Brian Greene how he reconciled physics’ static view of nature with the self-evidently dynamic experience of consciousness. His reply was that he “sees a psychiatrist,” that consciousness is capable of all sorts of illusions, and that the flow of time is just another example of the artifacts of consciousness.

While as detailed in the earlier section of this paper, I am the first to concede that our first-person reports can be fallible, as consciousness is capable of all sorts of illusions, it is hard for me to conceive of how consciousness could create an illusion of the flow of time, or the privileged present. There are several reasons why I am skeptical of this claim. First, just as matter must have extension in space in order to exist, so too it seems that consciousness must have extension in time. If consciousness had no “thickness” in time, then I simply do not understand how it could exist any more than an object could exist without some extension in space. Time is the dimension in which consciousness extends. Although the objective duration of the specious present ([James 1918](#)) may be rather modest ([Pöppel 1997](#)) without at least some extension in time I do not see how there can be any consciousness at all. Second, my experience is defined in terms of the flow of time and a privileged present; the stream of my consciousness is essentially a succession of “nows,” with the present always entailing the bridge between the past now and the future now. In a nutshell, from my first-person perspective I find the reality of the flow of time and the privileged present as compelling as the existence of physical reality itself (which also could in principle be an illusion, [Descartes 1641/1996](#)).

4.4 Reconciling first- and third-person perspectives of reality

Those who subscribe to a strict material reductionist perspective insist that when first-person experience suggests characteristics of reality

that are not readily handled by a third-person account, that those aspects must be rejected. From a strict materialist perspective, the seemingly privileged knowledge afforded by subjective experience, the flow of time, and the unique significance of the present all must be disregarded as illusions of consciousness. But herein lies the rub. The third-person perspective on reality is adequate as long as it provides constructs that correspond to the core aspects of the first-person perspective. However, when that perspective requires me to abandon absolutely fundamental aspects of my experience, then I am forced to question the assumptions that impose that requirement.

Whether we acknowledge it or not, all of us must discern for ourselves what aspects of existence to take as axiomatic. By definition, axioms cannot be empirically proven or logically deduced, rather they are self-evident truths that must be taken as givens. Perhaps the most fundamental of all such axioms is that physical reality exists; i.e., that I am not residing in a solipsistic mirage. Ultimately, while I grant the ontological reality of the physical world, in an important sense I am less epistemologically certain of it than I am of partaking in subjective experience. Ultimately, the only thing that I can know with absolute confidence is that I am currently enjoying a first-person experience (Descartes 1996). Physical reality could be a dream, I could be a brain in a vat or the matrix, indeed even my past could be an illusion, but there is simply no question but that I am currently having an experience. It might be an illusory experience⁴, but even an illusory experience is still experienced. Thus, although it is conceivable that physical reality could be an illusion, it is inconceivable (at least to me) that the occurrence of my subjective experience could be entirely baseless. This leads me to conclude that the existence of subjective experience and all premises that necessarily underpin its existence must be treated on equal ontological grounds to that of physical reality. Accordingly, if we grant subjective experience an ontological status equivalent to that of

objective reality then we must seriously question any characterization of objective reality that challenges the essential qualities of subjective reality. While much of our subjective experience may be an illusion, it is very difficult to see how the privileged vantage of subjective experience, the flow of time, or the unique status of the present could be such. To quote the philosopher David Ray Griffin (2007): “The reality of time is a more fundamental and stubborn fact than the alleged facts on which its denial is based” (p. 119).

A variety of approaches has been offered to accommodate the seeming limitations of a purely physical accounting of consciousness. Idealism (Berkeley 1878; Goswami 1993; Hoffman 2008) responds to the seemingly superior ontological status of subjective experience (i.e., its existence is more certain than an inferred external reality) by suggesting that if one must be reduced to the other, then it should be physical reality that is seen to be an outgrowth of subjectivity, rather than the other way round (as the material reductionists contend). Although difficult to refute, idealism (at least in the macro sense of conscious beings creating reality with consciousness) appears to discount the independent existence of a natural world, and thus seems at odds with a scientific vantage.

Another approach for reconciling the seemingly incommensurate existence of the subjective and objective is to pose that they both exist as two interacting yet distinct realms. This approach (substance dualism) was favored by Descartes, but it has a serious logical deficiency (at least as originally formulated): if two realms are truly incommensurate and distinct, then there seems to be no way for them to interact. To posit a “ghost in the machine” (Ryle 2009) is to assume that the ghost can affect the machine, which means that they share some common ground and therefore are not entirely distinct realms. This difficulty has proven a major problem for substance dualism (Armstrong 1999), although see Chalmers (2002) for arguments as to why the challenge of understanding the causal nexus between the mental and physical is not unlike similar issues of causality observed within the physical realm.

⁴ An illusory experience being defined as an experience that does not correspond to actual reality, such as a hallucination. Note that a philosophical zombie does not have an illusory experience of being conscious, it has no experience at all.

In my view, the seeming impasse between the third- and first-person perspectives of reality strongly suggests the existence of some other meta-perspective that can accommodate them both. Like the reversible images that can initially invoke one of two entirely opposed interpretations, but that can subsequently be reconciled from a vantage that recognizes the reality of both, (even if they cannot be both apprehended simultaneously) so too it seems there must be some meta-perspective for reconciling first- and third-person vantages on reality. In other words, it seems likely that there exists a higher order outlook that simultaneously acknowledges the manner in which neither perspective can simply be reduced to the other, yet still offers a mode of resolution. It is clearly easier to recognize the need for a meta-perspective than to identify precisely what such a view might be. Nevertheless, it seems a goal well worth pursuing.

Over the years, a number of scholars have tried their hand at envisioning a vantage that neither tries to reduce the subjective to the physical, nor the physical to the subjective, but rather conceives of some common ground or property that may be reflective of both. This approach, often referred to as neutral monism (Chalmers 2002; Feigl 1958; James 1904; Russell 1927), though with close affinities to dual aspect theories (e.g., Jackson 1982; Nagel 1986; Spinoza 1677/1985; Velmans 2009), attempts to identify a neutral realm of existence that can be alternately characterized as mental, physical, or neither.

The ever-changing present represents a core element of the common ground between subjectivity and objectivity that is invoked in several accountings of neutral monism. For William James, the neutral realm was the present:

The instant field of the present is at all times what I call the ‘pure’ experience. It is only virtually or potentially either object or subject as yet. For the time being, it is plain, unqualified actuality, or existence, a simple that. (1904, p. 23)

For Bertrand Russell, the neutral realm was the event: “Everything in the world is composed of

‘events.’... An ‘event,’ as I understand it ... is something occupying a small finite amount of space-time.” For Alfred North Whitehead (1929), the present also served as the nexus of conjunction between the objective and the subjective. In Whitehead’s panpsychic characterization of reality, the interface between first- and third-person perspectives occurs in the “creative advance” of the present in which time marches forward in a continual alternation among all elements of reality between subjective and objective states (for further discussions of Whitehead’s account, see Griffin 2007; Hunt 2011).

Information represents a second element that unites several efforts to find the neutral realm from which both subjectivity and objectivity arise. As Chalmers (1996) observes:

Perhaps, then, the intrinsic nature required to ground the information states is closely related to the intrinsic nature present in phenomenology. Perhaps one is even constitutive of the other. That way, we get away with a cheap and elegant ontology, and solve the two problems in a single blow. (pp. 304–305)

Sayre (1976) similarly argues that “the concept of information provides a primitive for the analysis of both the physical and the mental.” The notion that information somehow serves as the interface between the subjective and the objective is also a central component of Tononi’s (2008) recent suggestion that consciousness arises when matter produces “integrated information,” which is defined as “the amount of information generated by a complex of elements, above and beyond the information generated by its parts” (p. 216). The basic idea is that complex systems that integrate information, even potentially non-biological ones, will experience some minimal amount of consciousness: something it is like to be that system (see also Koch 2012, 2013).

In sum, although there is considerable variability in the manner in which scholars have conceptualized the common ground of reality from which both the objective and subjective emerge, two common elements are 1) that the

interface occurs within the ongoing march of the present, and 2) that it is constituted within the shared informational properties entailed in both objective and subjective states. A final shared aspect of many of these approaches is that subjectivity represents a fundamental attribute of the universe that either permeates all aspects of matter (panpsychism), or exists as a potentiality of matter that emerges when certain conditions are met (protopanpsychism; Chalmers 2002). Drawing on these general observations, I turn now to offering my own highly speculative conjectures regarding a meta-perspective on reality that may provide the shared foundation for first- and third-person perspectives.

4.5 The possibility of a subjective dimension of reality

Many scholars who posit that subjectivity is an essential aspect of reality argue that ultimately physics may need to be expanded to include constructs corresponding to subjective states. As the philosopher David Chalmers (1995) observed:

I propose that conscious experience be considered a fundamental feature, irreducible to anything more basic. ... In the 19th century it turned out that electromagnetic phenomena could not be explained in terms of previously known principles. As a consequence, scientists introduced electromagnetic charge as a new fundamental entity and studied the associated fundamental laws. Similar reasoning should be applied to consciousness. If existing fundamental theories cannot encompass it, then something new is required. (p. 96)

Eminent physicist Andrei Linde (1990) has also speculated that consciousness may some day be recognized as part of our understanding of physics:

Could it be that consciousness is an equally important part of the consistent picture of our world, despite the fact that so far one could safely ignore it in the description of

the well-studied physical processes? Will it not turn out, with the further development of science, that the study of the universe and the study of consciousness are inseparably linked, and that ultimate progress in the one will be impossible without progress in the other? (p. 27)

The critical question, of course, is: What in the physical universe might correspond to the arising of consciousness?

To recap, the physical realm as currently construed offers no place for subjective experience, the flow of time, or the uniqueness of the present. In order to bridge the gap between physical reality and subjectivity, scholars have posited the existence of a neutral realm that gives rise to both. Though varied in their emphasis, two elements have emerged as likely components of this neutral ground: the evolving present and information. Together these considerations suggest that a conjoined first-person/third-person meta-perspective will likely conceptualize subjectivity, the present, and the flow of time within an architecture that closely links information to an ever-changing now. Toward this end I offer the following conjecture: *consciousness arises via the changing informational states associated with an observer's movement through objective time relative to a currently unacknowledged dimension or dimensions of subjective time.*

Although speculative and highly underspecified, the above account has intuitive appeal. The sense of moving through time from one informational state to the next is clearly central to experience. Indeed it could well be said that it is the defining aspect of our existence. It is difficult to conceive of experience without invoking movement in time and change in informational state. Recall however that the current block universe portrayal of time provides no way to conceptualize moving through time, as movement in time would require change in time at a rate that could never be specified. As the Physicist Paul Davies observes:

But what meaning can be attached to the movement of time itself? Relative to what does it move? Whereas other types of motion relate one physical process to another, the

putative flow of time relates time to itself. Posing the simple question ‘How fast does time pass?’ exposes the absurdity of the very idea. The trivial answer ‘One second per second’ tells us nothing at all. (2002, p. 8)

subjective dimension (or dimensions) in which the observer moves relative to physical space-time (e.g., Smythies 2003). Noting the inability of current theories of physics to account for the flow of time or the existence of subjective experience, physicist Linde speculates that dimensions of consciousness may be required to provide the necessary degrees of freedom. Linde (2004) observes:

Thus to move in time requires movement in relationship to some dimension other than time itself. The postulation of an additional temporal dimension allows observers to change information states in objective time relative to subjective time. Indeed, it seems possible (and perhaps even a mathematical necessity) that in order to extend in and move through space-time (i.e., the block universe), there needs to be at least one additional dimension to provide the degree of freedom necessary to enable such movement (Schooler et al. 2011). In other words, if we accept the block universe model⁵ of reality, then in order to move through objective time, we have to move relative to something, and that something cannot itself be time because all time exists simultaneously in the block universe. A seemingly reasonable solution is to posit an additional dimension (or dimensions) of time. Although the postulation of additional dimensions of reality should not be taken lightly, it is not without precedent. In physics, string theory has postulated seven additional spatial dimensions beyond the three dimensions of space and one dimension of time that are customarily acknowledged (Greene 2004). If there can be multiple dimensions of space, then might there not also be additional dimensions of time? Indeed, some physicists have argued that an additional dimension of time might be very useful for conceptualizing various issues in physics (Bars et al. 1998). If the postulation of an additional dimension (or dimensions) of subjective time could also resolve the paradox of time and provide a realm for subjectivity, then surely that would also warrant its consideration as a possibility.

I am not the first to suggest that the failure of objective time (as it is currently conceptualized) to afford the flow of time or inner experience may require the postulation of an additional

Is it possible that consciousness, like space-time, has its own intrinsic degrees of freedom, and that neglecting these will lead to a description of the universe that is fundamentally incomplete? What if our perceptions are as real (or maybe, in a certain sense, are even more real) than material objects? What if my red, my blue, my pain, are really existing objects, not merely reflections of the really existing material world? Is it possible to introduce a ‘space of elements of consciousness’....? (p. 451)

I remain agnostic regarding precisely how many additional dimensions may be required in order to provide the degrees of freedom necessary for time to flow and consciousness to have extensions in the present. Indeed, I am not even committed to the notion that such a realm must necessarily be thought of as possessing all of the mathematical formalities of spatial dimensions. My point is simply that the current material reductionist model of reality has left no room for time to flow or now to exist. It is as if physics has built a pendulum clock but left no space for the pendulum to swing. In statistics, there always must be one more degree of freedom than the total number of subjects and conditions so as to leave the freedom for variables to vary. I believe that such degrees of freedom are similarly required to enable experience to flow through time.

A dynamic depiction of the value of adding a second temporal dimension is illustrated in the following three examples depicting a simple event of bottles breaking. The first (Figure 13; see video clip in its description) depicts the event as it would unfold from the first-person perspective, a dramatic shattering of initially intact colored bottles. The second example (Figure 14) transforms this event into a

⁵ Another possible way of reconciling the challenges of the flow of time and the present is to discard the notion of the block universe. While this vantage is the prevailing view in physics (Greene 2004), some have suggested that it needs revising (Hunt 2014; Smolin 2013).

block universe depiction in which objective time is spatialized, and each slice corresponds to a separate moment of the event. Notice that in the block universe representation there is no motion (and hence no video clip), and no singular frame (i.e., slice) corresponds to “now.” However, in the third example (Figure 15; see video clip in its description), an additional temporal dimension is introduced so that the observer can move through the block universe. Frame by frame a moving “now” marches through the block universe. By adding a second temporal dimension to the block universe, the dynamical experience of events unfolding is once again achieved.

A spatialized depiction of the notion of observers moving through subjective time relative to physical space and objective time is presented in Figures 16–18. As previously, noted, in the standard presentation of the block universe the three dimensions of space are, for graphical depiction, reduced down two dimensions (Figure 16). Here, in order to provide room to depict an additional dimension, physical space is further reduced to one dimension (Figure 17). Within this characterization, it is possible to see how the introduction of an additional dimension of subjective time (Figure 18) provides the necessary degree of freedom to enable an observer to move through time, as they can now move through physical time via a succession of moments in subjective time.



Figure 13: An event of breaking vases as it would be experienced from a first-person perspective. See http://open-mind.net/videomaterials/schooler_bootle_loaf5.mp4/view.

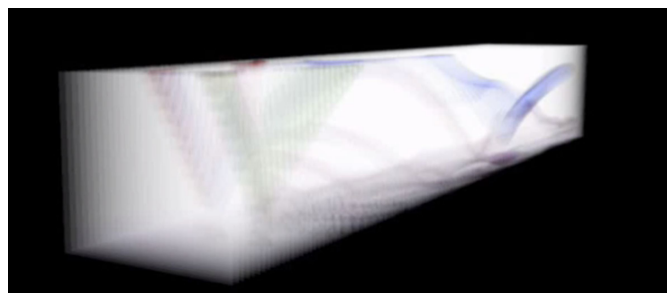


Figure 14: The breaking vases event is depicted as a block universe, with the temporal dimension spatialized, and each moment corresponding to a separate “slice.” Notice that there is no way to depict “now” and no way to move through it.

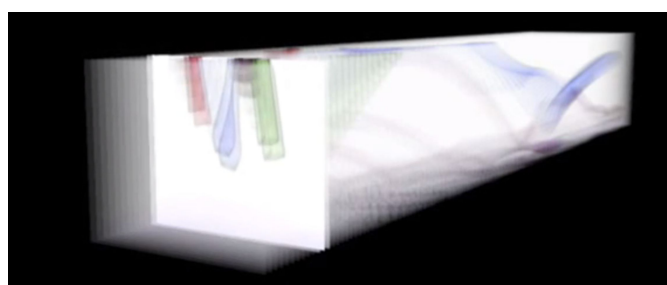


Figure 15: The breaking vases event is again depicted as a block universe, with the addition of a second temporal dimension. The moving present is represented as successive illuminated slices that progress from moment to moment through the block universe. Notice that witnessing movement through the block universe requires an additional dimension of time as the standard dimension of objective time is already dedicated to spatializing the block universe. See <http://open-mind.net/videomaterials/schooler-bottles-loaf-1.mp4/view>.

An interesting implication of this characterization is that observers can vary in the granularity (i.e., extent) of their moments. Notice how in Figure 17, the observer with the smaller spatial extent also occupies smaller successive moments in time.⁶ Intriguingly, there is evidence to support this view: recent findings suggest that smaller vertebrates may have a different “temporal grain size” relative to larger verteb-

⁶ Although subjective agents may move in subjective time relative to objective time in varying sized steps, it does not appear that there is necessarily a single temporal grain size for the processing of all sensory stimuli. Specifically, Pöppel (1997) finds that the duration of what constitutes a single moment (as assessed by temporal discrimination of successive sensory events) varies between sensory modalities. This observation seems potentially consistent with the suggestion that even within a single individual there may be multiple distinct conscious systems (Schooler et al. 2011; Zeki 2003) corresponding to different sensory levels and systems.

rates. Specifically, Healy et al. (2013) report a negative correlation between vertebrate size and the highest rate at which they can detect the flickering of a light (the flicker fusion rate). From the vantage of the current discussion, these findings suggest that the consciousness of smaller animals may move through subjective time relative to physical time at a faster rate than larger animals. This may be why it is so hard to swat a fly: from its vantage, we are moving in slow motion.

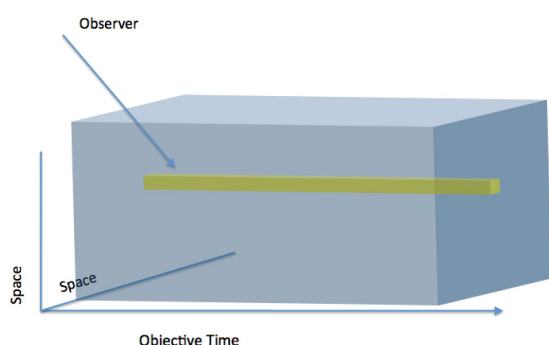


Figure 16: The observer depicted in the standard block universe with two dimensions of space. In the standard block universe, the observer is static and exists simultaneously in all locations. There is an insufficient number of degrees of freedom for the existence of a genuine now or movement in time.

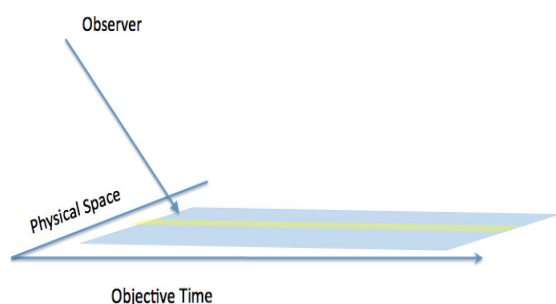


Figure 17: The observer depicted in a standard block universe with one dimension of space. As with the standard convention of depicting the block universe in two spatial dimensions instead of three, the reduction to one spatial dimension is useful for illustrative purposes.

A critical question that arises in postulating an additional subjective dimension (or dimensions) of time is: what are the properties of this dimension? I have left the answer to this question intentionally vague as I believe under-

specification leaves greater room to flesh out the rudimentary idea in various possible ways. With that said, it seems plausible that the subjective temporal dimension(s) could correspond to subjective informational states in the same way that objective informational states correspond to different moments of objective time. As noted, subjective informational states are aligned with but not identical to objective informational states (recall Mary, the color scientist). Moreover, current theories of neutral monism posit information as being one of the core potential interfaces between the objective and the subjective. Thus, characterizing subjective time as corresponding to distinct subjective informational states that are aligned with but not identical to objective informational states seems a promising characterization of the nexus between the objective and the subjective.

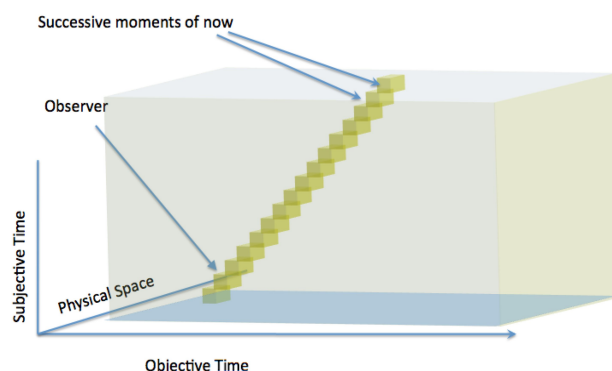


Figure 18: The observer depicted moving through a dynamic block universe with one dimension of physical space and the introduction of an additional subjective temporal dimension. In this model, there are a sufficient number of degrees of freedom to enable the observer to move in objective time relative to subjective time. The present can also be depicted as a series of moments extending in subjective time, objective time, and physical space.

A further potential benefit of the conjecture that experience emerges from movement in a subjective temporal dimension relative to objective time is that it provides a potential way of conceptualizing the nature of experience in the universe. Many scholars throughout history, and particularly those sympathetic to neutral monism, have articulated some type of panpsychic vision of nature, where all elements of

matter are seen as partaking in some rudimentary experience or proto-experience. Advocates of some version of panpsychism include [Spinoza \(1677/1985\)](#), [Leibniz \(1989\)](#), [James \(1909\)](#), [Bergson \(1896/1912\)](#), and [Whitehead \(1929\)](#). More recent adopters of this view include [Hameroff & Powell \(2009\)](#), [Chalmers \(1995\)](#), [Hunt \(2011, 2014\)](#), [Koch 2013](#), [Schooler et al. \(2011\)](#), [Skrbina \(2005\)](#), and [Strawson \(2008\)](#). The notion that the flow of time emerges by virtue of movement in a subjective temporal dimension relative to an objective one provides a potential way of conceptualizing how all of matter may partake in experience at varying levels of complexity. Accordingly, if experience emerges by movement through a dimension of subjective time relative to objective time, then it seems quite plausible that elements associated with all of matter may be on a shared trajectory through these two (or more) temporal dimensions, and thus may be enjoying some form of experience. In other words, if consciousness emerges from something as potentially ubiquitous as movement through an additional time dimension, then it seems plausible that all matter could enjoy some modicum of experience.

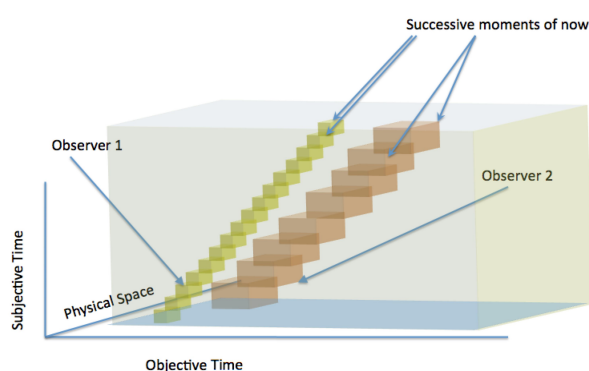


Figure 19: Two observers depicted moving through a dynamic block universe. Notice how this account enables varying temporal grain sizes between observers.

Although the present view provides a way of conceptualizing how matter might partake in at least some rudimentary form of experience, it need not suggest that all objects—collections of matter—are themselves sentient beings. To use [Nagel’s \(1974\)](#) terminology, there need be nothing “that it is like to be” a rock, for example.

Rather, the claim is that at some level, the constituent elements of a rock (and all other material objects) partake in at least some very rudimentary kind of experience, what the physicist/philosopher [Alfred North Whitehead \(1929\)](#) referred to as “actual entities”. In other words, according to the panpsychic tradition, matter is constituted of collections of individual elements each of which partake in some minimal experience. The subjective state of these individual experiential elements (or “actual entities”) is presumed to be extremely simple, and for the most part, when they combine, it is assumed that they form “mere aggregates” that do not entail a higher-order experience. However, under some circumstances, and in particular when present in certain organic structures, these simple actual entities may combine to form higher-order actual entities corresponding to the conscious agents that we typically acknowledge as such.

The notion of observers moving through objective time relative to a subjective temporal dimension may offer a possible direction toward solving the perennial “combination problem” of panpsychism, namely discerning how rudimentary proto-experiences of individual elements can combine to form the larger higher-order experiences that we enjoy ([Hunt 2011](#)). Accordingly, it seems possible that experience may correspond to oscillations in objective time relative to subjective time. As depicted in [Figures 18 and 19](#), I have speculated that observers may move in subjective time relative to objective time in discrete steps. The precise timing of these steps from one moment to the next could potentially provide the foundation for a unified experience among elements (i.e., an approach to the combination problem). When elements oscillate in synchrony (i.e., when they all jump from one moment in subjective time to the next), this may produce a unity of conscious experience. Nervous systems may provide an organizational structure that enables material elements to oscillate in synchrony and thereby produce larger, more organized fields of subjective experience. In this sense, the combination problem may be addressed by, and our holistic experience may result from, the common wavelength of oscilla-

tion through objective time relative to subjective time that constituent elements of a singular experience partake in. Put colloquially, each of us may have our own unique wavelength moving through subjective time relative to objective time.

Importantly, these speculations are presented as an example of the kind of meta-perspective that might enable an acknowledgement of the reality of both first- and third-person vantages. This is far from a formal model, and leaves much unspecified. For example, although I believe it could be possible to formalize the relationship between subjective time and informational states, this remains a major conjecture. Other elements of the framework, such as the notion that observers move in discrete steps in subjective and objective time, and that the pattern of oscillation may provide a way of addressing the “combination problem,” also are merely conjectures. I suspect that there are potentially a great variety of ways of conceptualizing how observers might move in a dimension of subjective time relative to objective time. My goal in attempting a rudimentary depiction of this notion is simply to fuel the conversation.⁷

Even if scientists resist the suggestion of an additional temporal dimension of reality, characterizing how experience can reside in a physical world will require explicating how observers move in physical time relative to changes in subjectively apprehended information. In other words, to be an observer in reality is arguably to reside in a now that corresponds

to a “location” within continually changing information states. Thus, conceptualizing the experience of the observer requires understanding how that observer moves between informational states over time. Given that the present prevailing view of physics does not afford the degrees of freedom to actually move in time, understanding how an observer changes informational states relative to time seems to require at a minimum the postulation of a virtual dimension of subjective time. Whether that dimension is given ontological status as a genuine aspect of reality depends on one’s perspective, but that of course is precisely the point.

For those who are willing to entertain the possibility of the kind of meta-perspective that I am envisioning, there are a number of possible ways forward. Perhaps, and most dramatically, it seems plausible that the existence of an additional temporal dimension may have empirical consequences. Although received with understandable skepticism, evidence continues to accumulate for precognition (i.e., that the mind is sensitive to events that have not yet occurred). There is a long tradition of research in this area (Honorton & Ferrari 1989). For example, Bem (2011) recently published a series of nine studies in a highly respected journal that seem to suggest evidence of genuine precognition and a subsequent meta-analysis of 90 additional findings appear to further substantiate these findings (Bem et al. 2014). Not surprisingly, these claims have been met with considerable skepticism (e.g., Ritchie et al. 2012; Wagenmakers et al. 2011). Given their profound challenge to our current scientific understanding of reality, claims of this sort will require studies that offer highly tangible evidence that cannot be attributed to artifact or statistical anomaly, e.g., taking advantage of people’s alleged precognitive capacities to make consistent future predictions of real world events, such as the future outcome of roulette wheel spins or the stock market (Franklin et al. in press). Nevertheless, the demonstration of robust findings of precognition might provide the type of data that could inform theories of how consciousness interfaces with time in a manner not currently considered in modern science.

⁷ Several years ago, I presented the idea that consciousness entails movement through a subjective dimension of time using the depiction in Figure 15 and illustrated at the site: <http://open-mind.net/videomaterials/schooler-bottles-loaf-1.mp4/view>. One of the attendees, Robert Forman (see his description of the event, Forman 2008), suggested that although he was intrigued by my depiction, that it did not square with his intuitions. In my model, the block universe is fixed and consciousness marches through it. He suggested that his intuition was the opposite: namely that the field of the observer remains fixed and time passes by, or changes within it. This alternative vantage in which time evolves through a fixed observer seems a worthy alternative perspective for conceptualizing the ever-changing now that may be closer to approximating several other neutral monist vantages (e.g., Whitehead 1929, and Hunt 2014). While I think this alternative viewpoint is worthy of consideration, I also think it is likely that the two vantages are logically equivalent—it is simply a question of which one is taken as the fixed frame of reference. Nevertheless the manner in which we construe the movement of time relative to the individual may have important psychological consequences (Casasanto & Boroditsky 2008).

Other approaches for fleshing out the kind of meta-perspective suggested here may include quantitative reconceptualization of existing findings. Although quantum theory is one of the most precisely predictive theories ever conceived, its explanation remains a mystery. In particular, the manner in which measurement seems to affect outcomes, and the theoretical relationship between measurement, consciousness, and the collapse of the wave function are not at all understood (Chalmers 2002). It seems possible that the postulation of an additional subjective dimension of time might lead to alternative ways of conceptualizing current formalism.⁸ Indeed it seems possible that once psychological constructs (such as a dimension of subjective time) are integrated with physical principles, that new psycho/physical laws of nature may emerge (Chalmers 2002; J. N. Schooler 2010). Alternatively, the notion that subjective experience emerges from movement through another dimension of time may resist empirical documentation, but may nevertheless remain a conjecture that appeals to some intuitions but not others.

Even if ultimately there is no conclusive ways of determining whether there exists an ad-

ditional subjective dimension of time this does not mean that the consideration or rejection of this view should be arbitrary. There are many judgments in life that rely on leanings that are not purely objective in nature. From ethics to art we routinely favor some views over others for reasons besides purely objective facts. Indeed the adoption or rejection of views close to those under discussion here are often based on subjective considerations. For example some physicists embrace string theory because of the elegance of its mathematics, whereas others reject it because there is no physical evidence to support it. Similarly there is great debate on how far down the phylogenetic scale we should postulate the existence of consciousness. Most of us have an opinion on this matter, but it remains entirely unclear whether there will ever be a purely objective way to resolve it. In the absence of objective evidence, our positions on these issues are far from arbitrary, rather they are based on the same sorts of sensibilities and intuitions that underpin many of our most heartfelt convictions.

In a final further effort to appeal to readers' intuitions, let me introduce one last metaphor for the meta-perspective I am striving for: consider the allegorical tale of *Flatland*, written by Edwin Abbott (1885) more than a century ago. Flatland depicts a two-dimensional world that is visited by a three-dimensional being (a sphere). The sphere takes a citizen of Flatland (a square) on a journey through the third dimension, offering the square a vantage on his reality that he never had before. The story of Flatland offers a number of useful lessons for the present discussion. First, it provides a powerful metaphor for thinking about the existence of additional dimensions of reality. Long preceding relativity theory, which treats time like a fourth dimension, or string theory, which currently posits the existence of up to seven additional spatial dimensions (Greene 2004), Abbott's tale introduces us to the concept of higher-order dimensions. Flatland describes how additional dimensions can be both embedded in and yet simultaneously transcend what we know. The parallels to consciousness are also striking: when the square is taken through the

⁸ Although not a mathematician it seems plausible to me that existing mathematical formalisms might be adopted to accommodate some of the present conjectures. Most speculatively, a quantitative characterization of additional dimensions of time might correspond in some manner to the many worlds account of quantum mechanics (Everett 1957) that postulates that every potential alternative outcome of quantum events entails a different branching parallel universe. It strikes me as possible that these so called "many worlds" could correspond to different coordinates in additional temporal dimensions. From this vantage, the block universe might be better conceived of as a block multiverse, with innumerable distinct temporal projections. Several multi-dimensional theories of objective time might (e.g., Bars et al. 1998; Craig & Weinstein 2008) also be relevant. Also potentially pertinent are various existing quantitative efforts to reconcile experience and physical matter. For example, the magnitude of a conscious observer's extension in subjective time might correspond to Tononi's (2008) quantitative assessment of Φ (pronounced "fi") which he characterizes as corresponding to the amount of integrated information that a conscious observer apprehends at any particular moment. Other potentially relevant formal approaches for reconciling consciousness with physical matter include: Hameroff & Penrose's (2014) efforts to explain how consciousness may "consists of a sequence of discrete events, each being a moment of 'objective reduction' (OR) of a quantum state" (p. 73), and Tegmark's (2014) suggestion that "consciousness can be understood as a state of matter, 'perceptronium', with distinctive information processing abilities" (p. 1). The relevance of these various approaches is highly speculative, and indeed given their disparities it is unlikely that they could be mutually accommodated. My point in mentioning them is simply to point the way towards some more formal approaches that might hold potential for advancing this discussion.

third dimension, he suddenly sees inside the objects of Flatland. Like consciousness, movement in an additional dimension in Flatland enables the perception of an inside where none could otherwise be possible. Like consciousness's relationship to reality, an additional dimension intersects with the lower dimensions and yet is distinct from them. And like the recognition of an additional dimension in Flatland, positing consciousness as moving through objective time relative to a dimension (or dimensions) of subjective time provides an example of a meta-perspective that potentially offers observers a new way of conceptualizing their relationship with physical reality. Although I make no claims as to having fleshed out this meta-perspective, it is my hope that my arguments have persuaded at least some readers that it is a vantage worth considering.

5 Summary and final conclusions

In this paper I used the thesis that perspective shifting can fundamentally alter how we conceive and evaluate evidence as the backdrop for exploring one of the most perennial and challenging of all perspectives shifts: namely, between the subjective first-person perspective that provides each of us with a unique window onto reality, and the objective third-person perspective that serves as the consensual foundation for science. My arguments were divided into three sections, which though admittedly distinct in their focus, all converge in attempting to elucidate a rapprochement between the subjective and objective perspectives on human experience.

In the [first](#) section I introduced the notion of perspective shifting in the context of classic reversible images. Here I argued that reversible images provide a context for conceptualizing how the very same situation can be understood from two very different perspectives that appear to produce seemingly irreconcilable accounts of their contents. However, once this juxtaposition is recognized, a meta-perspective emerges that enables the appreciation of both perspectives even if they cannot be apprehended simultaneously. The perspective shifting and meta-perspective that arise from reversible images

provide a metaphor for conceptualizing the tension between the first- and third-person perspective for understanding human experience. Both researcher and the field of science itself have been divided on whether to take perspectives on human nature that emphasize inner experience or external behaviors. While historically this has been a debate on which researchers have been forced to take sides, I argue that we should strive towards a meta-perspective in which the two vantages can inform one another.

In the [second](#) section I sought to show how the third-person perspective of objective science can elucidate our understanding of first-person experience. Towards this end, I introduced the distinction between having an experience (experiential consciousness) and one's explicit understanding of that experience (meta-awareness). Historically when researchers have sought to understand people's actual experience they have relied on people's self-reports about what they believe they were experiencing. This has led some to argue that it is impossible to gain insight into underlying experience. However, I argue that through triangulation between self-reports and behavioral and physiological measures, it is possible to make reasoned inferences about people's actual experience; identifying both situations in which meta-awareness overlooks experience (temporal dissociations of meta-awareness) and cases in which it distorts them (translations dissociations of meta-awareness). This framework was fleshed out within an extensive review of research on mind-wandering that, because of its inherently private nature, provides an ideal testing ground for developing a third-person science of first-person experience. By assessing the relationship between people's behavioral and physiological measures and self-report this review concludes that while people's self-reports of mind-wandering routinely correspond to genuinely experienced instances of this mental state, they nevertheless often fail to notice mind-wandering while it is occurring.

In the [final](#) and most speculative section of this paper, I turned the tables around. Instead of asking how third-person science clarifies first-person experience, I asked how first-person experience may inform third-person science. Here I ar-

gued that there are certain aspects of first-person experience that are so fundamental that they may reasonably serve as axioms of existence that any construal of physical reality must be able to accommodate. As detailed in the [prior](#) section it is clear that many aspects of experience may be illusory but several can reasonably be construed as unassailable, including: the occurrence of experience, the flow of time and the privileged present. Notably, current accounts of physical reality offer no way of accommodating these inherent aspects of first-person experience. This conflict between seemingly self-evident aspects of personal experience and current accounts of physical reality leads me to posit that, like the reversible images that can only be accommodated by recognizing a larger meta-perspective in which they both reside, so too there must exist some meta-perspective that can accommodate both objective scientific facts and personally experienced ones. Towards this end I introduced a highly speculative conjecture about the larger framework in which both objective and subjective perspectives might reside. Namely that consciousness involves a fundamental aspect of the universe that arises via the changing informational states associated with an observer's movement through objective time relative to a currently unacknowledged dimension or dimensions of subjective time. Although highly speculative, I offer this account as an example of the kind of meta-perspective that may simultaneously accommodate extant objective observations and certain aspects of subjective experience that I find as compelling as the existence of physical reality itself.

In my view, bridging the objective/subjective divide will require adopting a meta-perspective in which the two points of view are viewed as alternative vantages on an underpinning reality that corresponds to both but can be fully accommodated by neither alone. As I attempted to illustrate at the outset, it is quite possible to hold inaccurate or incomplete beliefs about one's experience, and third-person science can help to illuminate such errors. However, from my vantage there are certain elements of subjective experience that are as axiomatic as any aspect of the physical realm. Nevertheless, I recognize that not all will see it this way. Some will remain exclusively

fixed to the third-person perspective of objective science, while others will conceive of reality exclusively from their own personal first-person point of view. In conceptualizing this breadth of perspectives, it is important to remain mindful of an essential insight of Bayes' theorem of probability. Bayes' theorem states that in calculating the probability of something one must integrate new evidence with one's *a priori* probabilities. From a Bayesian perspective, for those who believe that something is impossible (i.e., infinitely unlikely) there is no amount of evidence or argument that should sway them. The ontological reality of first-person experience seems very much to fit in this category. My arguments on this point will likely remain wholly unpersuasive to those who cannot conceive of subjective experience as offering an epistemological authority that rivals science. However, for those open to the possibility that science will need to find a way to accommodate the reality of both the subjective and objective perspectives, I hope my discussion offers some glimmers as to what such a meta-perspective might be like.

Acknowledgements

The writing of this paper was possible with the support of grants from the Templeton Foundation, the Institute of Educational Studies, and the Fetzer Franklin Fund. I am grateful to numerous individuals who commented on earlier versions of this manuscript, including Ben Baird, Robert Bernstein, James Broadway, Ashley Brumett, Michael Franklin, Tam Hunt, Jack Loomis, Ben Mooneyham, Michael Mrazek, Brett Ouimette, John Protzko, Claire Zedelius, and two anonymous reviewers. This work has also benefited from conversations with David Chalmers, Daniel Dennett, Daniel Gilbert, Mark Laufer, Wolfgang Lukas, Merrill McSpadden, Brianna Morseth, Dawa Tarchin Phillips, Daniel Povinelli, Carmi Schooler, Lael Schooler, Nina Schooler, Rachel Schooler, Edward Slingerland, Jan Wallecezk, Dan Wegner, Timothy Wilson, Sid Zagri, and many others too numerous to mention. Though they aided in the project and/or influenced my thinking, none of these individuals or organizations should be viewed as endorsing the sentiments presented here.

References

- Abbott, E. A. (1885). *Flatland: a Romance of Many Dimensions*. Boston, MA: Roberts Brothers.
- Adams, H. E., Wright, L. W. & Lohr, B. A. (1996). Is homophobia associated with homosexual arousal? *Journal of Abnormal Psychology*, 105 (3), 440-440. [10.1037/0021-843X.105.3.440](https://doi.org/10.1037/0021-843X.105.3.440)
- Antrobus, J. S. (1999). Toward a neurocognitive processing model of imaginal thought. *At play in the fields, of consciousness: Essays in honor of Jerome L. Singer* (pp. 1-28). Mahwah, NJ: Erlbaum.
- Armstrong, D. M. (1999). *The mind-body problem: An opinionated introduction*. New York, NY: Perseus.
- Asendorpf, J. B. & Scherer, K. R. (1983). The discrepant repressor: differentiation between low anxiety, high anxiety, and repression of anxiety by autonomic-facial-verbal patterns of behavior. *Journal of Personality and Social Psychology*, 45 (6), 1334-1334. [10.1037//0022-3514.45.6.1334](https://doi.org/10.1037//0022-3514.45.6.1334)
- Baird, B., Smallwood, J., Fishman, D. J., Mrazek, M. D. & Schooler, J. W. (2013). Unnoticed intrusions: Dissociations of meta-consciousness in thought suppression. *Consciousness and Cognition*, 22 (3), 1003-1012. [10.1016/j.concog.2013.06.009](https://doi.org/10.1016/j.concog.2013.06.009)
- Baird, B., Smallwood, J., Lutz, A. & Schooler, J. W. (2014). The decoupled mind: Mind-wandering disrupts cortical phase-locking to perceptual events. *Journal of Cognitive Neuroscience*, 26 (11), 2596-2607. [10.1162/jocn_a_00656](https://doi.org/10.1162/jocn_a_00656)
- Baird, B., Mrazek, M. D., Philips, D. T. & Schooler, J. W. (in press). Domain-specific enhancement of meta-cognitive ability following meditation training. *Journal of Experimental Psychology*.
- Bars, I., Deliduman, C. & Andreev, O. (1998). Gauged duality, conformal symmetry, and spacetime with two times. *Physical Review*, D58 (066004). [10.1103/PhysRevD.58.066004](https://doi.org/10.1103/PhysRevD.58.066004)
- Bayne, T. (2015). Introspective insecurity. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100 (3), 407-425. [10.1037/a0021524](https://doi.org/10.1037/a0021524)
- Bem, D., Tressoldi, P. E., Rabeyron, T. & Duggan, M. (2014). *Feeling the future: A meta-analysis of 90 experiments on the anomalous anticipation of random future events*. Rochester, NY: Social Science Research Network.
- Bergson, H. (1912). *Matter and memory (original work published 1896)*. New York, NY: McMillan.
- Berkeley, G. (1878). *A treatise concerning the principles of human knowledge*. Philadelphia, PA: J.B. Lippincott & Company.
- Bloom, P. (2009). *Descartes' baby: How the science of child development explains what makes us human*. New York, NY: Basic Books.
- Carriere, J. S., Cheyne, J. A. & Smilek, D. (2008). Everyday attention lapses and memory failures: The affective consequences of mindlessness. *Consciousness and Cognition*, 17 (3), 835-847. [10.1016/j.concog.2007.04.008](https://doi.org/10.1016/j.concog.2007.04.008)
- Casasanto, D. & Boroditsky, L. (2008). Time in the mind: Using space to think about time. *Cognition*, 106 (2), 579-593. [10.1016/j.cognition.2007.03.004](https://doi.org/10.1016/j.cognition.2007.03.004)
- Chalmers, D. J. (1995). The puzzle of conscious experience. *Scientific American*, 273 (6), 80-86. [10.1038/scientificamerican0402-90sp](https://doi.org/10.1038/scientificamerican0402-90sp)
- (1996). *The conscious mind: In search of a fundamental theory*. New York; NY: Oxford University Press.
- (2002). Consciousness and its Place in nature. In D. Chalmers (Ed.) *Philosophy of mind: Classical and contemporary readings*. Oxford, UK: Oxford University Press.
- Cheyne, A. J., Solman, G. J., Carriere, J. S. & Smilek, D. (2009). Anatomy of an error: A bidirectional state model of task engagement/disengagement and attention-related errors. *Cognition*, 111 (1), 98-113. [10.1016/j.cognition.2008.12.009](https://doi.org/10.1016/j.cognition.2008.12.009)
- Christoff, K., Gordon, A. M., Smallwood, J., Smith, R. & Schooler, J. W. (2009). Experience sampling during fMRI reveals default network and executive system contributions to mind wandering. *Proceedings of the National Academy of Sciences of the United States of America*, 106 (21), 8719-8724. [10.1073/pnas.0900234106](https://doi.org/10.1073/pnas.0900234106)
- Churchland, P. S. (1989). *Neurophilosophy: Toward a unified science of the mind-brain*. Cambridge, MA: MIT press.
- Craig, W. & Weinstein, S. (2008). *On determinism and well-posedness in multiple time dimensions*. arXiv.org: 0812.0210.
- Crick, F. (1994). *The astonishing hypothesis*. New York, NY: MacMillan.
- Csikszentmihalyi, M. (1988). The flow experience and its significance for human psychology. In M. Csikszentmihalyi & I. S. Csikszentmihalyi (Eds.) *Optimal experience: Psychological studies of flow in consciousness* (pp. 15-35). Cambridge, UK: Cambridge University Press.

- Davies, P. (2002). That mysterious flow. *Scientific American*, 287 (3), 40-47. [10.1038/scientificamerican0206-6sp](https://doi.org/10.1038/scientificamerican0206-6sp)
- Dennett, D. C. (1993). *Consciousness explained*. London, UK: Penguin.
- (2003). Who's on first? Heterophenomenology explained. *Journal of Consciousness Studies*, 10 (9), 19-30.
- (2014). *Intuition pumps and other tools for thinking*. New York, NY: W.W. Norton & Company.
- Descartes, R. (1996). *Descartes: Meditations on first philosophy: With selections from the objections and replies (original work from 1641)*. Cambridge, UK: Cambridge University Press.
- Donchin, E. & Coles, M. G. (1988). Is the P300 component a manifestation of context updating? *Behavioral and Brain Sciences*, 11 (03), 357-374. [10.1017/S0140525X00058027](https://doi.org/10.1017/S0140525X00058027)
- Duncker, K. (1945). On problem solving. *Psychological Monographs*, 58 (5)
- Einstein, A. (2001). *Relativity: The special and the general theory. (Reprint of 1920 translation by Robert W. Lawson ed.)*. London: Routledge.
- Everett, H. (1957). 'Relative State' formulation of quantum mechanics. *Reviews of Modern Physics*, 29, 454-462.
- Farrin, L., Hull, L., Unwin, C., Wykes, T. & David, A. (2003). Effects of depressed mood on objective and subjective measures of attention. *The Journal of Neuropsychiatry and Clinical Neurosciences*, 15 (1), 98-104. [10.1176/appi.neuropsych.15.1.98](https://doi.org/10.1176/appi.neuropsych.15.1.98)
- Feigl, H. (1958). The 'mental' and the 'physical'. *Minnesota Studies in the Philosophy of Science*, 2, 370-497.
- Forman, R. K. (2008). A watershed event. *Journal of Consciousness Studies*, 15 (8), 110-115.
- Franklin, M. S., Baumgart, S. L. & Schooler, J. W. (in press). Future directions in precognition research: More research can bridge the gap between skeptics and proponents. *Frontiers in Psychology: Perception Science*.
- Gallagher, H. L. & Frith, C. D. (2003). Functional imaging of 'theory of mind'. *Trends in Cognitive Sciences*, 7 (2), 77-83. [10.1016/S1364-6613\(02\)00025-6](https://doi.org/10.1016/S1364-6613(02)00025-6)
- Giambra, L. M. (1995). A laboratory method for investigating influences on switching attention to task-unrelated imagery and thought. *Consciousness and Cognition*, 4 (1), 1-21. [10.1006/ccog.1995.1001](https://doi.org/10.1006/ccog.1995.1001)
- Goswami, A. (1993). *The self-aware universe: How consciousness creates the material world*. New York, NY: Putnam.
- Graziano, M. S. (2013). *Consciousness and the social brain*. New York, NY: Oxford University Press.
- Greene, B. (2004). *The fabric of the cosmos: Space, time, and the texture of reality*. New York, NY: A.A. Knopf.
- Griffin, D. R. (2007). *Whitehead's radically different post-modern philosophy: An argument for its contemporary relevance*. Albany, NY: State University of New York Press.
- Hameroff, S. & Penrose, R. (2014). Consciousness in the universe: A review of the 'Orch OR' theory. *Physics of Life Reviews*, 11, 39-78. [10.1016/j.phrev.2013.08.002](https://doi.org/10.1016/j.phrev.2013.08.002)
- Hameroff, S. & Powell, J. (2009). Embodied Prediction. In D. Skrbina (Ed.) *Mind that abides: Panpsychism in the new millennium*. Amsterdam, NL: Benjamins.
- Healy, K., McNally, L., Ruxton, G. D., Cooper, N. & Jackson, A. L. (2013). Metabolic rate and body size are linked with perception of temporal information. *Animal Behaviour*, 86 (4), 685-696.
- Hoffman, D. (2008). Conscious realism and the mind-body problem. *Mind and Matter*, 6 (1), 87-121.
- Hoffmann, B. & Dukas, H. (1972). *Albert Einstein, creator and rebel*. New York, NY: Viking.
- Honorton, C. & Ferrari, D. (1989). "Future telling": A meta-analysis of forced-choice precognition experiments, 1935-1987. *Journal of Parapsychology*, 53, 281-308.
- Hunt, T. (2011). Kicking the psychophysical laws into gear a new approach to the combination problem. *Journal of Consciousness Studies*, 18 (11-12), 96-134.
- (2014). *Eco, Ego, Eros: Essays on Philosophy, Spirituality and Science*. Aramis Press: Santa Barbara, CA.
- Hurlburt, R. T. & Heavey, C. L. (2001). Telling what we know: Describing inner experience. *Trends in Cognitive Sciences*, 5 (9), 400-403.
- Husserl, E. (1963). *Ideas: A general introduction to pure phenomenology. Trans. W. R. Boyce Gibson*. Collier Books: New York, NY.
- Jack, A. & Roepstorff, A. (2002). Introspection and cognitive brain mapping: From stimulus-response to script-report. *Trends in Cognitive Sciences*, 6 (8), 333-339. [10.1016/S1364-6613\(02\)01941-1](https://doi.org/10.1016/S1364-6613(02)01941-1)
- Jackson, F. (1982). Epiphenomenal qualia. *The Philosophical Quarterly*, 32 (127), 127-136.
- James, W. (1904). A world of pure experience. *Journal of Philosophy, Psychology and Scientific Methods*, 1, 477/533-491/543.
- (1909). *A pluralistic universe*. New York, NY: Longmans, Green, and Company.
- (1918). *The principles of psychology (Original work published 1890)*. New York, NY: Henry Holt and Company.

- Kam, J. W., Dao, E., Farley, J., Fitzpatrick, K., Smallwood, J., Schooler, J. W. & Handy, T. C. (2011). Slow fluctuations in attentional control of sensory cortex. *Journal of Cognitive Neuroscience*, 23 (2), 460-470. [10.1162/jocn.2010.21443](https://doi.org/10.1162/jocn.2010.21443)
- Katz, D. (1989). *The world of touch* (L. E. Krueger, Trans., original work from 1925). Hillsdale, NJ: Erlbaum.
- Klinger, E. (1999). Thought flow: Properties and mechanisms underlying shifts in content. In J. A. Singer & P. Salovey (Eds.) *At play in the fields of consciousness: Essays in honor of Jerome L. Singer* (pp. 29-50). Mahwah, NJ: Erlbaum.
- Koch, C. (2012). *Consciousness: Confessions of a romantic reductionist*. Cambridge, MA: MIT Press.
- (2013). Is consciousness universal? *Scientific American Mind*, 25 (1)
- LaBerge, S. P. (1980). Lucid dreaming as a learnable skill: A case study. *Perceptual and Motor Skills*, 51 (3f), 1039-1042. [10.2466/pms.1980.51.3f.1039](https://doi.org/10.2466/pms.1980.51.3f.1039)
- Lambie, J. A. & Marcel, A. J. (2002). Consciousness and the varieties of emotion experience: A theoretical framework. *Psychological Review*, 109 (2), 219-259. [10.1037/0033-295X.109.2.219](https://doi.org/10.1037/0033-295X.109.2.219)
- Leibniz, G. (1989). Monadology. In R. Ariew & D. Garber (Eds.) *G. W. Leibniz: Philosophical essays*. Indianapolis, IN: Hackett Publishing Company.
- Linde, A. D. (1990). *Particle physics and inflationary cosmology*. Chur, CH: Harwood Academic.
- (2004). Inflation, quantum cosmology, and the anthropic principle. In J. D. Barrow, P. C.W. Davies & C. L. Harper (Eds.) *Science and ultimate reality: Quantum theory, cosmology, and complexity*. Cambridge, UK: Cambridge University Press.
- Lutz, A. & Thompson, E. (2003). Neurophenomenology: Integrating subjective experience and brain dynamics in the neuroscience of consciousness. *Journal of Consciousness Studies*, 10 (9-10), 31-52.
- Manly, T., Robertson, I. H., Galloway, M. & Hawkins, K. (1999). The absent mind: Further investigations of sustained attention to response. *Neuropsychologia*, 37 (6), 661-670. [10.1016/S0028-3932\(98\)00127-4](https://doi.org/10.1016/S0028-3932(98)00127-4)
- Marks, I. M. (1987). *Fears, phobias, and rituals: Panic, anxiety, and their disorders*. New York, NY: Oxford University Press.
- Mason, M. F., Norton, M. I., Van Horn, J. D., Wegner, D. M., Grafton, S. T. & Macrae, C. N. (2007). Wandering minds: the default network and stimulus-independent thought. *Science*, 315 (5810), 393-395. [10.1126/science.1131295](https://doi.org/10.1126/science.1131295)
- McCaig, R. G., Dixon, M., Keramatian, K., Liu, I. & Christoff, K. (2011). Improved modulation of rostral prefrontal cortex using real-time fMRI training and meta-cognitive awareness. *NeuroImage*, 55 (3), 1298-1305. [10.1016/j.neuroimage.2010.12.016](https://doi.org/10.1016/j.neuroimage.2010.12.016)
- McGuire, P., Paulesu, E., Frackowiak, R. & Frith, C. (1996). Brain activity during stimulus independent thought. *Neuroreport*, 7 (13), 2095-2099. [10.1016/S0920-9964\(97\)82485-1](https://doi.org/10.1016/S0920-9964(97)82485-1)
- McKiernan, K. A., D'Angelo, B. R., Kaufman, J. N. & Binder, J. R. (2006). Interrupting the "stream of consciousness": An fMRI investigation. *NeuroImage*, 29 (4), 1185-1191. [10.1016/j.neuroimage.2005.09.030](https://doi.org/10.1016/j.neuroimage.2005.09.030)
- McVay, J. C. & Kane, M. J. (2009). Conducting the train of thought: Working memory capacity, goal neglect, and mind wandering in an executive-control task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35 (1), 196-196. [10.1037/a0014104](https://doi.org/10.1037/a0014104)
- Metzinger, T. (2004). *Being no one: The self-model theory of subjectivity*. Cambridge, MA: MIT Press.
- (2013). The myth of cognitive agency: Subpersonal thinking as a cyclically recurring loss of mental autonomy. *Frontiers in Psychology*, 4 (931). [10.3389/fpsyg.2013.00931](https://doi.org/10.3389/fpsyg.2013.00931)
- Mrazek, M. D., Smallwood, J. & Schooler, J. W. (2012). Mindfulness and mind-wandering: Finding convergence through opposing constructs. *Emotion*, 12 (3), 442-442. [10.1037/a0026678](https://doi.org/10.1037/a0026678)
- Mrazek, M. D., Smallwood, J., Franklin, M. S., Chin, J. M., Baird, B. & Schooler, J. W. (2012). The role of mind-wandering in measurements of general aptitude. *Journal of Experimental Psychology: General*, 141 (4), 788-788. [10.1037/a0027968](https://doi.org/10.1037/a0027968)
- Mrazek, M. D., Franklin, M. S., Phillips, D. T., Baird, B. & Schooler, J. W. (2013). Mindfulness training improves working memory capacity and GRE performance while reducing mind wandering. *Psychological Science*, 24 (5), 776-781. [10.1177/0956797612459659](https://doi.org/10.1177/0956797612459659)
- Murray, D. J. (1993). A perspective for viewing the history of psychophysics. *Behavioral and Brain Sciences*, 16 (1), 115-137. [10.1017/S0140525X00029277](https://doi.org/10.1017/S0140525X00029277)
- Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 4, 435-450.
- (1986). *The view from nowhere*. Oxford, UK: Oxford University Press.
- (2012). *Mind and cosmos: Why the materialist neo-Darwinian conception of nature is almost certainly false*. Oxford, UK: Oxford University Press.
- Nisbett, R. E. & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84 (3), 231-259.

- Pöppel, E. (1997). A hierarchical model of temporal perception. *Trends in Cognitive Sciences*, 1 (2), 56-61. [10.1016/S1364-6613\(97\)01008-5](https://doi.org/10.1016/S1364-6613(97)01008-5)
- Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A. & Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences of the United States of America*, 98 (2), 676-682. [10.1073/pnas.98.2.676](https://doi.org/10.1073/pnas.98.2.676)
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124 (3), 372-372.
- Reichle, E. D. (2006). Theories of the “eye-mind” link: Computational models of eye movement control during reading. *Cognitive Systems Research*, 7 (2-3). [10.1016/j.cogsys.2005.07.001](https://doi.org/10.1016/j.cogsys.2005.07.001)
- Reichle, E. D., Reineberg, A. E. & Schooler, J. W. (2010). Eye movements during mindless reading. *Psychological Science*, 21 (9), 1300-1310.
- Ritchie, S. J., Wiseman, R. & French, C. C. (2012). Failing the future: Three unsuccessful attempts to replicate Bem’s ‘Retroactive facilitation of recall’ effect. *PloS one*, 7 (3), e33423. [0.1371/journal.pone.0033423](https://doi.org/10.1371/journal.pone.0033423)
- Robertson, I. H., Manly, T., Andrade, J., Baddeley, B. T. & Yiend, J. (1997). Oops!: Performance correlates of everyday attentional failures in traumatic brain injured and normal subjects. *Neuropsychologia*, 35 (6), 747-758. [10.1016/S0028-3932\(97\)00015-8](https://doi.org/10.1016/S0028-3932(97)00015-8)
- Russell, B. (1927). *The analysis of matter*. London, UK: Kegan Paul.
- Ryle, G. (2009). *The concept of mind*. London, UK: Routledge.
- Sayette, M. A., Reichle, E. D. & Schooler, J. W. (2009). Lost in the sauce: The effects of alcohol on mind wandering. *Psychological Science*, 20 (6), 747-752. [10.1111/j.1467-9280.2009.02351.x](https://doi.org/10.1111/j.1467-9280.2009.02351.x)
- Sayette, M. A., Schooler, J. W. & Reichle, E. D. (2010). Out for a smoke the impact of cigarette craving on zoning out during reading. *Psychological Science*, 21 (1), 26-30. [10.1177/0956797609354059](https://doi.org/10.1177/0956797609354059)
- Sayre, K. (1976). *Cybernetics and the philosophy of mind*. Atlantic Highlands, NJ: Humanities Press.
- Schooler, J. W. (2002). Re-representing consciousness: Dissociations between experience and meta-consciousness. *Trends in Cognitive Sciences*, 6 (8), 339-344. [10.1016/S1364-6613\(02\)01949-6](https://doi.org/10.1016/S1364-6613(02)01949-6)
- (2010). *Mental inertia: Limited free will and determinism*. Santa Barbara: Paper submitted in fulfillment of UCSB Research Mentorship program, University of California.
- Schooler, J. W., Fallshore, M. & Fiore, S. (1994). Epilogue: Putting insight into perspective. In R. J. Sternberg & J. E. Davidson (Eds.) *The Nature of Insight*. Cambridge, MA: MIT Press.
- Schooler, J. W., Reichle, E. D. & Halpern, D. V. (2004). Zoning out while reading: Evidence for dissociations between experience and metaconsciousness. *Thinking and seeing: Visual metacognition in adults and children* (pp. 203-226). Cambridge, MA: MIT Press.
- Schooler, J. W., Smallwood, J., Christoff, K., Handy, T. C., Reichle, E. D. & Sayette, M. A. (2011). Meta-awareness, perceptual decoupling and the wandering mind. *Trends in Cognitive Science*, 15 (7), 319-326. [10.1016/j.tics.2011.05.006](https://doi.org/10.1016/j.tics.2011.05.006)
- Schooler, J. W., Hunt, T. & Schooler, J. N. (2011). Reconsidering the metaphysics of science from the inside out. *Neuroscience, Consciousness and Spirituality* (pp. 157-194). Berlin, GER: Springer.
- Schooler, J. W., Mrazek, M. D., Baird, B. & Winkielman, P. (2015). Minding the mind: The value of distinguishing among unconscious, conscious, and metaconscious processes. *APA handbook of personality and social psychology, Vol. 1. Attitudes and social cognition* (pp. 179-202). APA handbooks in psychology.
- Schooler, J. W., Mrazek, M. D., Baird, B. & Winkielman, P. (in press). Minding the mind: The value of distinguishing between unconscious, conscious, and meta-conscious processes. In P. Shaver & M. Mikulincer (Eds.) *APA Handbook of Personality and Social Psychology. Vol. 1: Attitudes and Social Cognition*. Washington, DC: APA Press.
- Schooler, J. W. & Mauss, I. B. (2010). To be happy and to know it: The experience and meta-awareness of pleasure. *Pleasures of the brain* (pp. 244-254). New York, NY: Oxford University Press.
- Schooler, J. W. & Melcher, J. (1995). The ineffability of insight. In S. M. Smith, T. B. Ward & R. A. Finke (Eds.) *The creative cognition approach* (pp. 97-134). Cambridge, MA: MIT Press.
- Schooler, J. W. & Schreiber, C. A. (2004). Experience, meta-consciousness, and the paradox of introspection. *Journal of Consciousness Studies*, 11 (7), 17-39.
- Schubert, T. W. & Semin, G. R. (2009). Embodiment as a unifying perspective for psychology. *European Journal of Social Psychology*, 39 (7), 1135-1141. [10.1002/wcs.55](https://doi.org/10.1002/wcs.55)
- Schultz, D. P. & Schultz, S. E. (1992). *A history of modern psychology*. New York, NY: Harcourt Brace.

- Singer, J. L. (1988). Sampling ongoing consciousness and emotional experience: Implications for health. In M. J. Horowitz (Ed.) *Psychodynamics and Cognition*. Chicago, IL: University of Chicago Press.
- Skrbina, D. (2005). *Panpsychism in the West*. Cambridge, MA: MIT Press.
- Smallwood, J. M., Baracala, S. F., Lowe, M. & Obonsawin, M. (2003). Task unrelated thought whilst encoding information. *Consciousness and Cognition*, 12 (3), 452-484. [10.1016/S1053-8100\(03\)00018-7](https://doi.org/10.1016/S1053-8100(03)00018-7)
- Smallwood, J., Davies, J. B., Heim, D., Finnigan, F., Sudberry, M., O'Connor, R. & Obonsawin, M. (2004). Subjective experience and the attentional lapse: Task engagement and disengagement during sustained attention. *Consciousness and Cognition*, 13 (4), 657-690. [10.1016/j.concog.2004.06.003](https://doi.org/10.1016/j.concog.2004.06.003)
- Smallwood, J., O'Connor, R. C., Sudberry, M. V., Haskell, C. & Ballantyne, C. (2004). The consequences of encoding information on the maintenance of internally generated images and thoughts: The role of meaning complexes. *Consciousness and Cognition*, 13 (4), 789-820. [10.1016/j.concog.2004.07.004](https://doi.org/10.1016/j.concog.2004.07.004)
- Smallwood, J., Fishman, D. J. & Schooler, J. W. (2007). Counting the cost of an absent mind: Mind wandering as an underrecognized influence on educational performance. *Psychonomic Bulletin & Review*, 14 (2), 230-236. [10.3758/BF03194057](https://doi.org/10.3758/BF03194057)
- Smallwood, J., McSpadden, M. & Schooler, J. W. (2007). The lights are on but no one's home: Meta-awareness and the decoupling of attention when the mind wanders. *Psychonomic Bulletin & Review*, 14 (3), 527-533. [10.1016/j.tics.2011.05.006](https://doi.org/10.1016/j.tics.2011.05.006)
- Smallwood, J., O'Connor, R. C., Sudbery, M. V. & Obonsawin, M. (2007). Mind-wandering and dysphoria. *Cognition and Emotion*, 21 (4), 816-842. [10.1080/02699930600911531](https://doi.org/10.1080/02699930600911531)
- Smallwood, J., McSpadden, M., Luus, B. & Schooler, J. (2008). Segmenting the stream of consciousness: The psychological correlates of temporal structures in the time series data of a continuous performance task. *Brain and Cognition*, 66 (1), 50-56. [10.1016/j.bandc.2007.05.004](https://doi.org/10.1016/j.bandc.2007.05.004)
- Smallwood, J., McSpadden, M. & Schooler, J. W. (2008). When attention matters: The curious incident of the wandering mind. *Memory & Cognition*, 36 (6), 1144-1150. [10.3758/MC.36.6.1144](https://doi.org/10.3758/MC.36.6.1144)
- Smallwood, J., Beach, E., Schooler, J. W. & Handy, T. C. (2008). Going AWOL in the brain: Mind wandering reduces cortical analysis of external events. *Journal of Cognitive Neuroscience*, 20 (3), 458-469.
- Smolin, L. (2013). *Time reborn: From the crisis in physics to the future of the universe*. Houghton Mifflin Harcourt: Boston, MA.
- Smythies, J. (2003). Space, time and consciousness. *Journal of Consciousness Studies*, 10 (3), 47-56.
- Spinoza, B. (1985). Ethics (Ed. & Trans., original work published 1677). In E. Curley (Ed.) *The collected works of Spinoza (Vol. I)*. Princeton, NJ: Princeton University Press.
- Spreng, R. N., Stevens, W. D., Chamberlain, J. P., Gilmore, A. W. & Schacter, D. L. (2010). Default network activity, coupled with the frontoparietal control network, supports goal-directed cognition. *NeuroImage*, 53 (1), 303-317. [10.1016/j.neuroimage.2010.06.016](https://doi.org/10.1016/j.neuroimage.2010.06.016)
- Strawson, G. (2008). *Real materialism and other essays*. Oxford, UK: Oxford University Press.
- Tegmark, M. (2014). Consciousness as a state of matter. *arXiv:1401.1219v2*
- Tong, F., Meng, M. & Blake, R. (2006). Neural bases of binocular rivalry. *Trends in Cognitive Sciences*, 10 (11), 502-511. [10.1016/j.tics.2006.09.003](https://doi.org/10.1016/j.tics.2006.09.003)
- Tononi, G. (2008). Consciousness as integrated information: A provisional manifesto. *The Biological Bulletin*, 215 (3), 216-242.
- Velmans, M. (2009). *Understanding consciousness*. London, UK: Routledge.
- von Goethe, J. W. (1867). *Faust: A dramatic poem*. London, UK: Hamilton, Adams, and Company.
- Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Sing, M. & von der Heydt, R. (2012). A century of gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization. *Psychological Bulletin*, 138 (6), 1172-1172. [10.1037/a0029333](https://doi.org/10.1037/a0029333)
- Wagenmakers, E. J., Wetzels, R., Borsboom, D. & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi. *Journal of Personality and Social Psychology*, 100 (3), 426-432. [10.1037/a0022790](https://doi.org/10.1037/a0022790)
- Wallace, A. (2000). *The taboo of subjectivity: Toward a new science of consciousness*. Oxford, UK: Oxford University Press.
- Weissman, D., Roberts, K., Visscher, K. & Woldorff, M. (2006). The neural bases of momentary lapses in attention. *Nature Neuroscience*, 9 (7), 971-978. [10.1038/nn1727](https://doi.org/10.1038/nn1727)
- Whitehead, A. N. (1929). *Process and reality: An essay in cosmology*. Cambridge, UK: Cambridge University Press.

- Wigner, E. & Margenau, H. (1967). Remarks on the mind body question, in symmetries and reflections, scientific essays. *American Journal of Physics*, 35 (12), 1169-1170.
- Wilber, K. (1998). *The marriage of sense and soul: Integrating science and religion*. New York, NY: Broadway Books.
- Wilson, T. D. (2003). Knowing when to ask: Introspection and the adaptive unconscious. *Journal of Consciousness Studies*, 10 (9), 131-140.
- Wilson, T. D., Lisle, D. J., Schooler, J. W., Hodges, S. D., Klaaren, K. J. & LaFleur, S. J. (1993). Introspecting about reasons can reduce post-choice satisfaction. *Personality and Social Psychology Bulletin*, 19 (3), 331-331. [10.1177/0146167293193010](https://doi.org/10.1177/0146167293193010)
- Wilson, T. D. & Schooler, J. W. (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology*, 60 (2), 181-181. [10.1037//0022-3514.60.2.181](https://doi.org/10.1037//0022-3514.60.2.181)
- Zedelius, C. M., Franklin, M. S., Smallwood, J., McSpadden, M., Reichle, E. D. & Schooler, J. W. (2014). Unnoticed nonsense: Mind wandering can prevent people from realizing that they are reading gibberish (manuscript under review).
- Zeki, S. (2003). The disunity of consciousness. *Trends in Cognitive Sciences*, 7 (5), 214-218. [10.1016/j.tics.2011.11.016](https://doi.org/10.1016/j.tics.2011.11.016)

Bridging the Gap

A Commentary on Jonathan Schooler

Verena Gottschling

In my commentary on this rich paper, I will focus on the methodological approach proposed by Schooler. The main goal of this commentary is to introduce an improved and more detailed interpretation of Schooler's distinction between experiential consciousness and meta-awareness. I will address four issues. After summarizing Schooler's main ideas, I will discuss some general problems regarding the proposed distinction between experiential consciousness and meta-awareness. I will relate the distinction to the more general debate. I then discuss some conceptual claims to which Schooler seems to be committed to making, and show how they relate to one another. I point to some tension between them. As I will argue, the central issue has to do with the underspecified notion of "reflection". Different kinds of reflection are required for Schooler's "pure experience" and for meta-awareness. I will try to get a better grasp on the author's underlying position by discussing the main empirical evidence motivating the account, namely mind-wandering, in section two. I argue that the evidence does not support the distinction as introduced, but does give us some insight into the complexity of the required meta-cognitive processes. I will suggest some conceptual changes in the underlying framework, which I believe make the main project stronger and help to avoid some of the problems we have encountered. Specifically, I want to introduce a taxonomy of different kinds of reflection and show which kinds of reflections might required both for Schooler's "pure experience" and for his meta-awareness. In the third section, I turn to the author's main claim, which is the existence of a new meta-perspective. According to Schooler, this is the central proposal of his paper, and it follows from his initial perceptual-perspective-shifting analogy and the distinction he proposes. Schooler claims that the meta-perspective helps us to overcome the limitations of both perspectives: the first person perspective and the third person perspective. In effect, by introducing the meta-perspective we can bridge the gap between self-reported experiences and observable behavior, and get a completely new perspective on the mind-body problem. As I will argue, this ontological element is relatively independent of the rest of his methodological project. Moreover, it is an unnecessary strategic move.

Keywords

Accessibility | Cognitive | Consciousness | Higher-order accounts | Phenomenal | Reportability | Stream of consciousness

1 Introduction

Starting from perceptual perspective shifting, Schooler focuses on the gap between self-reported experiences (the first-person perspective) and observable behavior (the third-person perspective). So consciousness versus self-awareness of being in a certain state, and the relationship of both of these to observable behavior is at the heart of the project. The main goal of the tar-

get paper is to introduce a new methodology for studying conscious versus unconscious states and processes. Although this is a very rich paper, we are not given too much information about the conceptual framework and the way in which Schooler's proposal relates to the contemporary philosophical debate about consciousness, reportability, and accessibility. This aspect

Commentator

Verena Gottschling

vgott@yorku.ca

York University

Toronto, ON, Canada

Target Author

Jonathan Schooler

jonathan.schooler@psych.ucsb.edu

University of California

Santa Barbara, CA, U.S.A.

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität

Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University

Melbourne, Australia

will be my focus: the relationship between philosophical theories of consciousness and Schooler's account.

Schooler's ([this collection](#)) project uses the combined strategy of self-reports, observable behavior, and physiological measurements of the body: a "trust but verify" (p. 8) approach to reports of subjects' experience. He is interested in:

the relationship between people's belief about their experience and empirical indices of their *underlying mental states*. [...] Moreover, the theory of the intermittent and imperfect *nature of meta-awareness as a re-representation of experience* [...] provides a scaffold for conceptualizing the situations in which *beliefs and underlying experience converge and diverge*. (p. 19, emphasis added)

Though it sounds at the beginning as if Schooler is making a claim about internal states in general in general, it quickly becomes clear that he indeed makes a claim about the personal-level, or conscious internal states. In so doing, he transitions from internal states to a certain kind of internal state—a conscious one. Later in the paper we find similar transitions: first we find a statement that can be interpreted as talking about all internal states, or verbally reportable knowledge of one's states, but then he immediately makes a statement about the underlying experience. For example, he informs us that in mind-wandering we can "identify situations in which all evidence suggests people are routinely lacking in their current *knowledge of their on-going mental states*" ([Schooler this collection](#), p. 19, emphasis added). A little later we find a statement about experience, thus knowledge or beliefs about *conscious* states:

In short, a strong case can be made for the value of using 3rd person science to inform not only our understanding of people's beliefs about their experience, but also to discern when those beliefs are likely to be accurate and when they may be inaccurate or incomplete. ([Schooler this collection](#), p. 20, emphasis added)

A similar transition from a statement about internal states to a statement about conscious internal states, which as a result can be reported, can also found slightly earlier:

by using various reasonable markers of people's internal states we have been able to examine the conditions under which people's reports are more or less likely to be aligned with their experience. ([Schooler this collection](#), p. 19, emphasis added)

To summarize, it seems that "what is going on in someone's mind", in Schooler's terminology, refers to the conscious mind. His approach locates him in a group of thinkers¹ who challenge the notion of accurate reportability, or who challenge access as the main criterion for conscious experience. There is a very active contemporary dispute between defenders of what have been dubbed cognitive accounts of consciousness and proponents of non-cognitive accounts ([Overgaard & Grünbaum 2011](#)). Opponents of cognitive approaches associate consciousness with cognitive functions like controlled processing, working memory, selective attention, or some network of different cognitive processes.² Because of this association, these functions can be used to study consciousness from a third-person perspective. In contrast, non-cognitive approaches assume that consciousness cannot be operationalized in terms of cognitive function. Consequently, these accounts dissociate consciousness from cognitive capacities. Which leaves us (typically) with just subjective criteria as acceptable for studying consciousness. Obviously Schooler's account is an example of a cognitive approach. In my opinion, this general dispute cannot be resolved by empirical evidence because neither of these approaches can be empirically falsified, or at least the empirical evidence can in principle be explained both ways—in essence we have a clash of intuitions, and the evidence can be interpreted as supporting opposing views.³ However, the approach one favors

1 See for example [Seth et al. \(2005\)](#), who presented a proposal close in spirit.

2 See [Overgaard & Grünbaum \(2011\)](#); [Block \(2011\)](#); [Cohen & Dennett \(2011\)](#); [Kouider et al. \(2010\)](#).

3 See the debate about alternative explanations of the findings of atypical perceptual conditions (for example of the Sperling paradigm) in the references above.

will obviously determine one's criteria of consciousness, the experimental methodology used, and, consequently, one's findings. Nonetheless, I do not want to go too much into this very wide dispute, partly because I think it would be rather fruitless.⁴ So for the purposes of this commentary, I will focus on issues *within* cognitive approaches alongside Schooler's cognitive account. But the objections against cognitive accounts of consciousness in general are issues that Schooler, given his introduction of a cognitive methodological approach for studying consciousness, potentially needs to address.

By using mind-wandering as his main example, Schooler then proposes a list of criteria that—so the idea goes—might help us to get a better grasp on the conscious experience, and not just conscious states to which we attend or states of which we are meta-aware. This underlying conceptual distinction turns out to be essential for Schooler's overall project.

One way of interpreting Schooler's account is to see it as a combination of a number of claims, which is evident in the quote above. He himself, right after introducing the distinction, argues that the two cases come apart in mindreading, and the fact that “people routinely shift perspective (from simply experiencing to attempting to re-represent their experience to themselves) provides the foundation for a framework of scientifically investigating first person perspective” (Schooler [this collection](#), p. 9). The implicit *main argument of the paper* can be reconstructed in the following way:

(1) Schooler introduces a conceptual distinction between *experience* and *meta-awareness* as a re-representation of experience.

(2) He then presents empirical evidence that this conceptual distinction corresponds to reality, in mind-wandering and other cases.

(3) He then uses this evidence to suggest a general list of testable features for those interested in the empirical investigation of conscious-

ness. The last issue is particularly important: in effect, Schooler suggests replacing the classical testable criterion for consciousness, (oral) reportability, or accessibility to introspection, by several criteria, which are testable and available from the third-person perspective.

(4) He claims that this gives us a principled new way of reconciling the tension between the first- and third-person perspective by introducing a higher meta-perspective, an ontological claim; in essence, this meta-perspective allows for a new strategy to solve the mind-body problem. We are promised the above-mentioned new “framework for scientifically investigating first-person experience” (Schooler ([this collection](#), p. 9) resulting from the analogy of perspective shifting.

2 The revised view

There is much more in the target paper than I have mentioned here. For the purposes of this commentary, I will focus on four issues related to the general issue of consciousness, which then result in the presentation of a revised version of the author's account. Now that I have summarized what I take to be the author's most important ideas, I will discuss some general problems the underlying distinction seems to bring with it. This section receives my main attention. I will try to localize the distinction within theories of consciousness. I then discuss some underlying conceptual claims to which Schooler is committed to making, and show how they relate to one another. I will point out that there is serious tension between them. In the second section, I will discuss in more detail the main empirical evidence that motivates the account—mind-wandering—and introduce the proposed criteria. My epistemic goals in the commentary are, first, to determine the exact relationship between the initial distinction, the evidence presented, and the proposed list of criteria. Second, to discuss of how we should evaluate certain criteria, and what they tell us about underlying concepts of meta-awareness, access, and reflection. Third, to gain some insight into the relationship between one's position regarding the mind-body problem and the suggestion

⁴ By this I mean that the evidence does not allow us to rule out the whole class of cognitive versus non-cognitive accounts. I think however, that certain accounts within these classes are vulnerable to evidence; for example, explicit accessibility accounts (Prinz 2012) seem to have a lot less room to maneuver. But as a debate between cognitive versus non-cognitive accounts, the possibilities for interesting general insights seem limited.

the author draws from his perceptual perspective shifting analogy. According to Schooler, this is the central proposal of his paper; he claims the existence of a new-meta-perspective, which helps to overcome the limitations of both perspectives and thereby solves the mind-body problem. As I shall argue, this element is relatively independent from the rest of the project. Moreover, I think it weakens the main project.

As a positive contribution, I will suggest some conceptual changes of the underlying framework. The changes I will suggest include giving up some claims and revising others. I think these changes make the main project, which I take to be a methodological strategy for studying consciousness, stronger. They also help to avoid some problems we encountered in the discussion of the main argument. I also suggest a finer-grained specification of different kinds of reflection and taking stock. This will help to give us a better understanding of meta-cognition in general as well as of consciousness and awareness of being in a certain state as distinct phenomena. I take this to be a driving idea in Schooler's initial distinction.

3 The category of “conscious but unaccessed” states

Traditionally, we find a distinction in the literature between two categories: on the one hand conscious experiences, states, and processes to which subjects have access, and on the other hand unconscious processes to which they do not have access (Cohen & Dennett 2011). According to this general picture, access to these states and processes then includes in many cases accurate reportability, which is the reason why reportability, or accessibility to introspection, is central to any judgment about conscious states. But access can also be understood more broadly: not all access is conscious itself, and not all access results in behavioral or verbal reportability.

In general, if we have a conscious state and a corresponding unconscious state, there are two possibilities for how the two can differ.⁵

⁵ Of course hybrids are possible, so we might have combinations of functional differences and differences in content. I take Tye (1995) to defend such an account.

The first option is that the representational content of a state determines the experience, at least in part, so that both states differ in content. My conscious belief that my partner is cheating on me has a different representational content than the corresponding unconscious belief. These accounts are first-order accounts. The second option is that the states have identical representational content, but there is a difference in kind in the way in which they are embedded in the system—in philosophical jargon, the functional role that each state plays differs. According to this position, my conscious and unconscious suspicious beliefs that my partner is cheating on me are two states with the same content—expressed in the *that*-clause—but the conscious belief causes different internal states and different behavior to my unconscious belief. For example, in the conscious case, I will have the conscious thought that he is not treating me respectfully, and I might verbally confront him right away; in the second, unconscious case, neither of these activities will happen.

The first option is consistent with the standard view of what determines a difference in experience. However, it has a disadvantage: we cannot explain why the two states “correspond” unless there is some significant semantic overlap between them. The functional role view has the advantage that it can explain the similarity between the two states, but the disadvantage that we need an explanation of what exactly it is that makes a state conscious, and we have to show *why* this difference results in a difference in experience.

Schooler seems to opt for the content or representational view. Picking up Dennett's idea⁶ that people can be inaccurate about their own mental going-ons and internal states, Schooler concludes that, at least in some situations, external observers can have better insight into a subject's experience than the subject themselves (p. 8). However, as we saw in the quotes above, Schooler seems to interpret the internal states in question as conscious internal states.

This is consistent with the idea that the access to internal states changes the content of

⁶ See Schooler (this collection), p. 8.

the state, i.e., the content view: accessing a state changes the content of the state. Since the content determines the experience, the experience of a non-accessed and an accessed state differ. Understood this way, Schooler's criteria give us opportunities to know *better* than the subject himself what he consciously experiences. Access and the reports of subjects about their experience, and the experience itself can come apart. If this is right, it would be unexpected and not what the commonsense understanding of conscious states predicts. As for the first aspect, Schooler believes that mind-wandering gives us an empirical case, where accessing (in the sense of attending to) a process or state changes that very state.

3.1 The general distinction between conscious experience and meta-awareness

I will start with a discussion of the motivation for the distinction (see p. 3), and some general problems we seem to invite if we accept this distinction. Schooler, and with him others, presuppose that conscious experience and accessibility can come apart; moreover, there is an experience *before* it is accessed. In other words, we postulate a third category, besides conscious and unconscious states: there are now “conscious but not accessed” states. These thoughts seem to be in line with other considerations in this debate, which propose a new category of phenomenal consciousness with no access (Block 2011; Lamme 2003).

Schooler distinguishes between simply “having experiences”, which he calls that *experiential consciousness*, and explicitly “taking stock” or re-representing this experience, which he calls *meta-awareness* or *meta-consciousness* (Schooler 2002, p. 339). Meta-consciousness then, is “defined as the intermittent explicit re-representation of the contents of consciousness” (2002, p. 339), while a later he says it is “knowing that one is having that experience” (2002, p. 339). So meta-awareness is about a certain kind of access.

Because we can clearly distinguish both, mind-wandering seems an excellent empirical

candidate for the study of consciousness. At one point we notice our mind-wandering; but *what* we notice, the mind-wandering itself, occurs earlier. In the meta-aware case, we re-represent the former state; in order to do this, we access it by re-representing it, and we “take stock”. Then the subject becomes meta-aware of the state, and we know that we are in this state, but this very process changes the content. Our experience of mind-wandering is different once we become meta-aware that we are mind-wandering.

But this seems conceptually puzzling. Access and (verbal) reportability are clearly not the same, such that missing (verbal) reportability cannot not be equated with general lack of access, especially at the subpersonal level. With knowledge, reflection, re-representation and meta-awareness, as well as meta-consciousness, we get additional and differing concepts. First, often “knowledge” is used as something that is itself conscious. Is the idea that we are aware only of the mind-wandering, or also of our knowledge that we are mind-wandering? The author alternates between both phrases. But both claims differ. I can be aware of an experience without being aware of my knowledge that I have this experience. The latter includes a meta-level of a different kind. While the first contains a meta-process regarding the experience, the second is a meta-process referring to a propositional state, knowledge, *of* the experience. As a result I am aware of being in the state and not just of the experience. Moreover, reflection is a vague term. How exactly do we reflect on a state, process, or content of a state? What exactly does this entail? So the question is: what is meta-awareness and what distinguishes it from simple awareness? Finally, re-representation is mentioned, yet another concept used to characterize meta-awareness. Without further explanation, re-reflection seems a very broad and vague concept that would include all kinds of re-represented contents. Do most of these occur unconsciously, as certain kinds of functional accounts, higher-order accounts, predict (Jackendoff 1987; Rosenthal 2005)? How is something re-represented? How exactly

do the representation and the re-representation relate to one another?

The question of which types of neural processes might be sufficient for awareness is highly controversial in current debate, as is whether there can be any awareness of a state without access (see the exchange between [Fahrenfort & Lamme 2012](#) and [Cohen & Dennett 2011, 2012](#)). Relatedly, the status of local recurrences is debated. Block and Lamme argue that there are perceptual cases in which subjects do not attend to a stimulus (in change blindness, inattention blindness, and attentional blink) and as a result are not able to report the presence of the stimulus. They might nonetheless be phenomenally conscious of the stimulus because it induces local recurrence in perceptual brain regions. As a result, a subject's reports are not to be trusted in all cases: subjects could be conscious of stimuli even when they themselves deny it. This sounds very close in spirit to Schooler's idea. However, Schooler doesn't tell us how his account, and pure mind-wandering versus meta-awareness of mind-wandering, relates to this debate.

Despite these unclear aspects, the underlying intuitive idea is clear: Schooler wants to distinguish phenomenally-conscious experience from a meta-level of consciousness, in the literature also referred to as meta-awareness, and sometimes as reflective awareness, reflexivity, or reflexive consciousness. But what exactly characterizes this meta-level remains unclear. We are simply not told, the used concepts seem vague, and, without further explanation, underspecified. But, of course, this does not imply that the main idea is not helpful, or that it is not possible to specify them.

However, Schooler seems to sympathize with Cohen and Dennett, so I take it that he thinks (like them), that awareness differs from behavioral reportability. However, Cohen and Dennett explicitly state that they do not see many reasons to think such conscious information exists before it is accessed ([Cohen & Dennett 2012](#), p. 140). So they reject the very option, the third category, that Schooler wants to postulate. There seems to be a sharp tension between Schooler's distinction and his agree-

ment with Cohen and Dennett's general approach: Whereas Cohen and Dennett argue that theories postulating inaccessible conscious states are intrinsically off-limits to investigation, Schooler not only defends an account along those lines, but also argues that his account gives us a solution strategy to overcome the tension between the first- and third-person. Obviously, there is a need for conceptual clarification of this highly original idea.

However, I think we can learn a few interesting things from this. First, we can rule out a very general understanding of reflection or meta-cognitive processes. Most theorists agree that part of what it is to be in a conscious state is to have a unified perspective on the world. So the possibility of distinguishing between me and the world, or a self, or some kind of self-consciousness is required as an indispensable part of conscious experiences of many kinds. One way of describing this is to say that experience includes some kind of categorization. In other words, it is a kind of meta-cognition on this highest and most general level. At least, we as humans keep track of this interdependence of action and perception/experience at the personal level. To mention a classical example, it seems very hard to experience pain if one doesn't classify something as painful, or without seeing it as painful *for me*. Indeed, some kind of evaluation, conscious or not, seems to be required for something to classify as pain; just as, in order to see something visually as a cow, we have to classify or categorize it as a cow ([Dretske 1993](#)).

At first glance, an account like Schooler's cannot allow for this because the standard view requires meta-cognition for conscious experience. Experience is cognitively penetrable, such that knowledge about categories influences how we experience an object. In contrast, Schooler distinguishes both, and wants to allow for experience before (any?) meta-level involved. At least he talks sometimes as if meta-cognition in general is the issue when it comes to meta-awareness of mind-wandering. When he talks about theory of mind and the areas involved in meta-cognition ([Schooler this collection](#), p. 17), he suspects that because certain meta-cognitive

processes and mind-wandering occupying both engage the same systems, specially the dorsal ACC and the anterior PFC, this might explain why it is so hard to catch oneself mind-wandering, i.e., to gain meta-awareness of mind-wandering. However, he notices that identity of brain regions does not imply a causal relationship, and that further research is necessary.

However, it would be hasty to conclude that Schooler cannot concede that meta-cognition can be involved in experience on his account. Though he talks frequently as if the issue were meta-cognition in general, he is not committed to excluding *any kind of* meta-cognitive process. But what is needed is a differentiation between different kinds of meta-reflection or re-representation. Schooler needs to address the question of whether we see the same kind of meta-cognitive processes in different kinds of experiences, and how exactly this changes the experience. Interpreted this way, only a certain *kind* of meta-reflection or meta-cognition might establish meta-awareness. As I will show, this move avoids a number of other problems.

We know that experience depends on background knowledge, and that our knowledge and our classification processes change our experience in many cases. This seems to be the case not just in mind-wandering, Schooler's favorite example, but also in many other cases. What matters is not just how I classify a state or process; many other internal states and contextual factors influence experience. Let's assume that I am a big fan of Baroque music, but cannot stand twelve-tone music. I happen to blunder into a concert with music by Penderecki, and of course do not like what I hear. Simply by gazing at the program and learning that I am listening to Penderecki's Saint Luke Passion, which uses references to motives by Johann Sebastian Bach and is in a sense a homage to a well-known Bach piece, how I experience this piece of music might change. Chances are that I am still not able to hear the references to Bach and the coded references to passages in Lucas in the middle of all the dense tone clusters. But my belief that it is a homage to my beloved Bach will change my experience in general. Other states, beliefs, and emotions in-

fluence my auditory experience and make it, in this case, somehow more enjoyable.⁷ It is also well known that crossmodal influences change experience: one's taste experience changes with conflicting visual experience. So a pure strawberry juice tastes less like strawberry to us if it is colored blue, even if the juice itself is not altered.⁸ How we experience a certain wine depends on knowledge about price, how famous the winery is, and many situational aspects. In these examples, the real question seems to be how exactly our experience changes, and how do particular internal and external factors contribute to the change. What changes in how we re-represent, and how fundamental is this change? And what is meant by these terms?⁹

So Schooler's meta-awareness can come in many forms. "Meta-cognition" includes a broad range of phenomena. What they have in common is that subjects have some insight into their own cognitive functioning. It is not clear to me that it is an all-or-nothing affair between pure experience and meta-awareness or reflection. So a specification of what exactly is meant by meta-awareness, re-reflection, and access seems necessary. We also need to answer the question of how the two categorically differing states differ in content, and which *exact kinds of meta-processes* are relevant. "Reflection" and "re-representation" are notoriously vague terms. Some kind of reflection at least seem indispens-

7 Bayne & Montague (2011) provide a nice overview of the complex cognitive phenomenology debate in his introduction to his volume. One might think that other contents causally influence the phenomenology of a state. A second option would be that "what it is likeness" is not a useful conceptual distinction at all (Lycan 1996, p. 77; Papineau 2002, p. 227). A third option would be that there are several meanings of "what its likeness"—indeed, in the literature different distinctions have been suggested. I will go into more detail in a later section, when I introduce elements of an improved taxonomy.

8 See further discussion in Grush (this collection).

9 Regarding visual perception Siegel (2005) has argued that that learning to recognize an object can change the way that it looks—in the phenomenal sense of "look", which is taken to imply that the cognitive components of such states are necessary for explaining the change in phenomenal character. In contrast, one could argue that the phenomenology does change, but the change can be explained in sensory terms instead of in terms of cognitive components. Either a subject's concepts do not directly constitute the subject's phenomenal states, such that they can have a causal influence on their phenomenology (Carruthers & Veillet 2011), or the contrast between both is the result of differences in the way that one processes the information within the sensory system (Tye & Wright 2011). For my purposes here, what matters most is *that* the phenomenology differs, and that we need an explanation for it.

able for a state to be conscious. But that doesn't mean the distinction above is not justifiable. We just need to determine and specify the kind of reflection and/or re-representation. I will make some suggestions later in this paper (see p. 15).

To be fair, while Schooler does not distinguish between different kinds of reflections, he indirectly assumes that there are differences. But in his view the phenomenon dictates what the criterion for introspective awareness is. He distinguishes classification under the concept of "taking stock": "there are some mental states (e.g., mind-wandering) for which the crucial bottleneck in people's introspective awareness stems not from their capacity to classify the experience, but rather from the fact that people only intermittently take stock of what is going on in their own minds" (Schooler [this collection](#), p. 8).

This obviously implies that for other phenomena the crucial difference does stem from their capacity to classify an experience. As a result, we in effect have different criteria for introspective awareness and for mind-wandering and visual perception. I believe a more promising route is to allow for dimensions of reflection and complexity of experience along multiple dimensions, but to try to find as uniform criteria as possible. The experience and phenomenology in cases of thought and sensory states (broadly construed) might be different.¹⁰ But some properties or property clusters have to bind instances of introspective or meta-awareness together.¹¹ Otherwise, what would justify classifying them as the same, if both the phenomenon and the properties associated with the phenomenon differ? We would just be talking about different things. I have already ruled out two kind of meta-cognitive processes the author cannot use for a more detailed characterization of the difference between conscious states and meta-aware states: categorization under concepts is one kind of meta-cognitive reflection that itself is unconscious, but necessary for con-

scious experience. Distinguishing between self and world is another dimension of reflection, at the highest level, that seems necessary. Meta-cognition always requires representational use (of some kind), because within it we find monitoring of cognitive affordances. But there are several *ways* in which this monitoring can take place. As I argue below, meta-cognition, the ability to monitor and control one's own cognition, and the ability to attribute mental states to oneself and others can occur in different ways; and both the self-other distinction, and self-awareness can occur in a number of ways.

3.2 Meta-cognitive accounts of consciousness: Content vs. function

A core idea in the target paper is the claim that there is a difference between an experience and an experience one is aware of having. Both states are experienced, but the idea seems to be that reflection could potentially change an experience in a certain way, because it focuses on the content of the intentional formerly un-reflected state. Interpreted this way, Schooler seems to defend the content view, though I do not think he is committed to it. He doesn't explicitly subscribe to it, but it seems implicit in what he says when he talks about the content of states and frequently switches back and forth between content talk and talk of experience. He seems to think that these are related. And he doesn't say much about the functional role that the states in question play in other states, or how cognitive processes use them—something one would expect if he held the functional view. So it is tempting to interpret him as having the view that content determines experience (Block 2005). For example, in writing that there are "some situations in which observers might have better knowledge about a person's mental state than does the person in question" (Schooler ([this collection](#), p. 8), what he must mean is that observers have better insight into the *content* of people's states. A little later, he claims, regarding misrepresentations, "while in the process of re-representing, one omits, distorts or otherwise misrepresents one's mental state to oneself and/or others" (Schooler [this collection](#),

¹⁰ As the complex debate about the possibility of a phenomenology of thought suggests.

¹¹ At the very least we would need to insist that there is a family of co-occurring properties playing an explanatory role within theories (Boyd 1999).

p. 10). Again, what we misrepresent is obviously the *content of the state*.¹² If he has a content view, than his view is that (at least in some cases) I have an experience first, and then, when I reflect on it, that very process changes the content of the initial intentional state. That then is the reason why the experience differs between mind-wandering as “purely experienced”, and mind-wandering experienced with awareness. The phenomenon of mind-wandering indeed introspectively changes after we reflect upon it, and become aware that we are mind-wandering.

But I think there is a larger issue here. Interpreted this way, it is tempting to judge that accounts claiming that what makes a state a conscious state is its functional role are inconsistent with Schooler’s account. Again, I think this would be too hasty. Let me explain. Assuming a representational theory of phenomenal consciousness,¹³ there are accounts that provided in purely first-order terms and accounts that implicate higher-order cognition of one sort or another (see below) with conscious experience. If we accept Schooler’s distinction, a state is conscious before we are aware of it, or know that we are in this state, and, when we become aware of it, this changes the state, or its content, to be more precise, as Schooler seems to suggest. Thus, Schooler seems to defend a first-order account, namely an account in which it is claimed that the consciousness of a state is partly (or entirely) determined by its representational content, or sometimes the format of its representational content, not primarily at first the function it plays (Byrne 2001; Dretske 1993; Kriegel 2009).

In the class of functional¹⁴ accounts we find a great range of different accounts, including second-order accounts, accessibility accounts (Prinz 2012), and global workspace accounts (Baars 1988). Many of these are close in spirit to Dennett’s. Though they differ, they have one thing in common: it is a certain functional relationship the states in question have to other states or within the system, which makes these states conscious states.

Second-order accounts, for example, would claim that what makes a state a conscious state is that the state is (or is disposed to be, in some versions) the object of a higher-order representation of a certain sort. This state is a meta-level state, a mental state directed at another mental state. Higher-order accounts differ on how exactly this higher-order representation is characterized and what the exact relationship between both states is. In some versions the higher-order representation is a higher-order thought (Rosenthal 1986, 2005), in others a higher order-order perceptual or experiential state (Lycan 1996), yet other versions see the higher-order state as dispositional (Carruthers 2000). There are also differences concerning the question of whether the higher-order state should be understood as entirely distinct from its target state (Rosenthal), or whether the higher-order thought is better viewed as intrinsic to the target state, which would imply that we have a complex conscious state with parts. There exist different versions of the intrinsic view, which all have in common the idea that instead of a separate higher-order state there is a global meta-representation within a complex brain state (Gennaro 1996; Van Gulick 2000; Metzinger 1995). For the purposes of this commentary, I will focus on Rosenthal’s higher-order thought theory, but my considerations generalize to many of the higher-order accounts. The existence of the higher-order state and the right connection between both (one is the object of the other) makes the lower level one a conscious state. The higher-level state, however, is itself unconscious, unless there exists a third-level state—the existence of which would result in awareness of being in a conscious state. In effect, the existence of a certain kind of meta-cognition is what makes the lower level state a conscious state, or even a state that we are aware of being in. In this framework, Schooler’s meta-awareness would require a third-order state.

Accessibility accounts, for example that of Jesse Prinz’ (2012), would claim that attention is both necessary and sufficient for states to be conscious. In global availability accounts¹⁵ it is

¹² See also, for example Schooler (this collection), pp. 16-17.

¹³ For the purposes of this commentary I neglect biological state theories.

¹⁴ On a very broad reading of “functional”.

¹⁵ Initially introduced by Baars (1988, also 1996). More modern proponents would be, for example, Dehaene et al. (2006).

claimed that the functional role is the global availability, or the workspace. The idea is that there is competition among neural coalitions; the winning coalitions are the conscious ones. There are a lot of similarities between higher-order theories and the neuronal global-workspace theory, but we should not see them as theories of the same type. According to the neuronal global-workspace theory, a state is conscious due to the global availability of its content, whereas higher-order theories see a state's being conscious as "consisting of one's being aware of oneself as being in that state" (Rosenthal 2012, p. 1433). If one interprets Rosenthal's reference to "oneself" as Metzinger's phenomenal self-model (2003), then a higher-order theory requires the integration of an individual state in a coherent representation or inner model of oneself, in contrast to a global-workspace theory, in which all that is required is availability of the content. Both aspects, the kind of meta-representation (the number of higher-order steps) and a certain identification of the original state as *my* state are dissociable, and they are examples of what I mean by different dimensions of reflection.

I think Schooler's account stands in natural alliance with both kinds of accounts, in contrast to what one might initially think. It is the vagueness of the term "meta-awareness" that is causing this unjustified reluctance. For example, higher-order thought accounts seem a natural way to specify what Schooler might have in mind when he talks about meta-aware states. According to Rosenthal, there can be unconscious pain states, if these are accompanied by the thought that I am in pain, I am experiencing pain, but the thought itself is unconscious. Only if there is a third-order state, the thought that I have the thought of being in pain, am I aware that I think that I am in pain. To me, this sounds close to Schooler's meta-awareness of taking "stock of our ongoing experience and re-present[ing] it to ourselves" (this collection, p. 8). However, there is an important difference: for Rosenthal there are only conscious and unconscious states; the presence of the third-order state gives us what Schooler might call meta-awareness. However, Rosenthal

denies the very possibility Schooler claims exists, that one can be in a conscious state but not aware of it. "No mental state is conscious if the individual that is in that state is in no way aware of it" (Rosenthal 2012, p. 1425). Due to the existence of a third-order state with the right content, we get introspective awareness of a conscious state: a third-order awareness that makes one aware of the second-order awareness. Rosenthal expects such cases, in which we "are aware of focusing attentively on that state" (2012, p. 1427), to be rare. It seems to me that there is a natural fit between Schooler's meta-aware states, in which we know that we are having a certain experience and Rosenthal's introspective awareness of a conscious state. In Rosenthal's framework, meta-awareness necessarily requires a third-order representation.

In addition, Schooler's suspicion that "meta-awareness appears to be associated with rhythms of attentional flux" (this collection, p. 17) relates nicely to accessibility accounts.¹⁶ But as I will claim in the next section, global availability accounts stand in another obvious alliance with Schooler. Again, it seems that it all depends upon our understanding and further specification of "reflection" or the "meta" in Schooler's meta-awareness. Is reflection itself necessarily a conscious process? Is it a thought, or just any kind of representation for the purposes of monitoring one's own cognition or an explicit higher-order classification? Unfortunately, Schooler does not describe his meta-awareness in more detail.

It seems to me that we should concede that some kind of "reflection" might be required for something to be an experience. This leaves still plenty of room to specify different kinds of reflections, some of which might constitute more than awareness, namely meta-awareness. This becomes the real question. Is this reflection itself unconscious or even necessarily conscious? Is it a re-representation of some kind? If that is the case, what kind of re-representation is required? Schooler's meta-awareness might require a rather demanding kind of reflection, and the

¹⁶ However, in the end accessibility accounts will not be Schooler's best bet—after all, I interpreted him above as agreeing that access and awareness differ.

relationship Rosenthal describes seems a good candidate. But perhaps what we have instead of a simple dichotomy between pure experience and meta-awareness is a full spectrum of dimensions of meta-representation. Then the question is, what are the dimensions of reflection required for Schooler's "pure experience" and those for meta-awareness, and which other reflections are there? This search for a proper taxonomy of "reflection" seems the most pressing need. It will hence be my main focus, and I will suggest some building blocks for such a taxonomy (p. 15). Rosenthal's introspective awareness of a conscious state as an possibility for characterizing Schooler's meta-awareness will be one element of this.

3.3 A general concern for scientific practice and a conceptual worry

This brings us to another and more problematic issue. I find the general line of thought behind a rigid distinction between pure experience and meta-awareness of this experience problematic. First, it presupposes that we accept the distinction between access-consciousness and phenomenal consciousness—a distinction not everybody (to say the least) is happy to accept.¹⁷ Second, and more fundamentally, such a new category would have to be motivated. How do we distinguish "conscious processes, which are not accessed" from unconscious activity? Are they *de facto* not explicitly re-represented, or is it impossible to re-represent them? What does it then mean to say that something is "conscious"? One might suspect that this new concept of "conscious" is not compatible with our common-sense intuitive understanding of the term. Moreover, the stronger reading of Schooler's position might invite further problems. If we claim that access to a state would necessarily change the status of its content (or the content itself), it would be impossible to address whether it was of a phenomenal or unconscious nature prior to this conscious access. If

such an "observer-effect" exists, it could potentially render the whole issue completely immune to scientific investigation (Kouider et al. 2012).

Another open question is how Schooler's account relates to others that seem close in spirit. Dehaene et al. (2006) have presented a more modern and updated version of Freud's concept of preconscious activity. They introduce a proposal with a carefully defended taxonomy of three categories: subliminal, preconscious, and conscious activity. According to Dehaene and Changeux's workspace model developed a little later, dominant neural coalitions involving the workspace are *accessed*. In contrast, existing other weaker activations in the workspace, such as a connection that *could be activated*, for example by a shift of attention, are only *accessible*. Processes that are potentially accessible, but are not accessed at the moment because of sufficient top-down attentional amplification, are "preconscious" phenomenal conscious processes in Dehaene et al.'s terminology (2006, pp. 206-207). I am not sure whether what Schooler is proposing is another version of Dehaene et al.'s "preconscious" phenomenal consciousness. This is consistent with what he writes. In debates on the third category of phenomenally conscious but not accessed states, their distinction between cognitive access and cognitive accessibility is often used to defend the possibility of the aforementioned third category (see for example Block 2011). My own suggestion is related, although I will suggest more closely specifying different kinds of access (see p. 11) and multiple levels of representation, instead of just distinguishing between accessibility and access.

I think Schooler's account would profit from directly relating his terminology to other concepts already in use in the debate. However, there are problems looming: Dehaene et al. defend a version of a functional account, which Schooler seems to explicitly reject when he seemingly advocates a first-order account. But if Dehaene's taxonomy is *not* what the author has in mind, what is the difference between the Schooler's phenomenally conscious but unaccessed activities and Dehaene's preconscious activities?

¹⁷ However, one might be able to resist the distinction between access-consciousness and phenomenal consciousness and at the same time allow for Schooler's distinction between experienced consciousness and meta-awareness if one claims that access is *not* what characterizes the meta-level in Schooler's meta-awareness.

Let us take stock. I have argued so far for three closely related points. The basic distinction between being experientially conscious of a state and being meta-conscious of being in a state needs further conceptual clarification. Moreover, the combination of a first-order account of consciousness (the content view) and this very distinction might not be the optimal strategy. In fact, a functional or hybrid account seems to provide a more natural strategic alliance for Schooler's main project. Finally, it seems there is no strict dichotomy between experiential consciousness and meta-awareness; we rather face a difference in many dimensions. From my perspective, both higher-order accounts as well as global workspace accounts might be helpful regarding this issue. They connect nicely with Schooler's main project, and would help to clarify his basic distinction. But we might very well end up with a more complex understanding of different meta-cognitive dimensions and differentiations instead of a simple conceptual dichotomy. This is what I will provide later in this paper. In order to do this we need to take a closer look at Schooler's second step; his argument that his conceptual distinction is something we find in cognitive capacities.

4 Mind-wandering—and noticing it. The bundle of criteria

On the basis of the former considerations he presents, Schooler argues that in many capacities we actually find a difference between being in a certain state and noticing that one is in a state (meta-awareness). So he moves onto his second claim, the claim that his conceptual distinction is empirically supported (see p. 3). According to Schooler, there are two forms of dissociations, *temporal dissociations* on the one hand and *translation dissociations* (misinterpretations) on the other. Let me begin with temporal dissociations. Examples of temporal dissociations are mind-wandering vs. noticing one's mind-wandering, but also mindless behaviors, suppressed thoughts, and unwanted emotions. Schooler mostly uses mind-wandering, however, characterized as situations, in which

we “lose track of the contents of our own minds” (Schooler [this collection](#), p. 9). This is the starting point for the introduction of Schooler's new “framework for scientifically investigating first-person experiences” (Schooler [this collection](#), p. 9).

I find this focus on mind-wandering a little puzzling, because I am not sure why this is an example supporting the general claim that the content of individual states changes in the specific intentional states. Why is it an individual intentional state that changes? Mind-wandering (at least intuitively) seems to be a complex process, and involves a number of states. In mind-wandering the issue is creature consciousness, not the experience or phenomenal character of an individual state, i.e., state-consciousness. Mind-wandering is about a train of thoughts, often accompanied by emotions, and autobiographical memories. In mind-wandering, we mostly think about issues related to our own life. For example we consider our “to-do” lists for today, what to have for dinner, our relationship to people close to us, telephone calls we need to make, and even our next lecture. At least the phenomenal character we experience during mind-wandering seems to include these the associated sensory states—broadly construed to include feelings of emotions, images, moods—which have a distinctive “phenomenal character” or “what it's likeness”. But the stream of consciousness also contains episodes of conscious thought.¹⁸ If we use this standard understanding of mind-wandering, it would rather be a bundle of thoughts, associations, or states, in other words a number of many more or less related thoughts, emotions, or other states and processes, not all of them necessarily fully specified in terms of content. And if so, it is not necessarily the content of individual states that changes—we seem to have multiple options for characterizing what changes once we

¹⁸ Schooler (2013) gives a good overview of the performance costs associated with mind-wandering (including reading comprehension, model building, and impairment of the veto-option to automatized responses) and suggests that mind-wandering may represent a pure failure of cognitive control. For this reason it is so useful to study consciousness. He argues that mind-wandering offers little benefit, though it might have a positive role in topics related to autobiographic episodes and information, for example in autobiographical planning and creative problem-solving.

are aware that we are mind-wandering. An alternative interpretation would be that the network of associated elements might change, or even the kind of associations involved. For a conceptual analysis, whether one should include these autobiographic sensory states in the phenomenon itself or just say the “train of thoughts in mind-wandering” causes them, is unclear. But it will determine how we analyze the experience of mind-wandering and the meta-awareness of mind-wandering, and its implications for theories of consciousness. There is also evidence that it has different functions and might itself be a heterogenic phenomenon (Northoff 2014, especially chap. 26; Metzinger 2013). For example, it is not clear whether mind-wandering is the same as day-dreaming, and if not, what the differences are.

Moreover, it is controversial whether thoughts even *have* a phenomenal character, and if so, how to analyze it (Bayne & Montague 2011). The orthodox view is that conscious thoughts themselves do not have a distinctive “phenomenal character”. They are either considered conscious without phenomenal character, or it is conceded that conscious thoughts might possess phenomenal character, but only in virtue of the sensory states with which they are associated (for example Braddon-Mitchell & Jackson 2007; Carruthers 2005; Nelkin 1989; Tyne 1995). However, recently, a number of authors introduced views according to which conscious thoughts themselves possess a “distinctive” phenomenology, but the phenomenal character differs from sensory states (Siewert 1998; Pitt 2004; Robinson 2005; Prinz 2004).

So there are a lot of further issues to consider, for a project like Schooler’s; we need to analyze the experience of mind-wandering and contrast it with meta-awareness or reflective experience in mind-wandering. However, Schooler gives some other examples for temporal dissociations, which can more obviously be explained in terms of individual states we do not notice or misinterpret. He doesn’t go into detail, but has mentioned mindless behaviors, suppressed thoughts, and unwanted emotions. The idea seems to be that we are not aware of an individual unwanted emotion, or a thought that

causes behavior. However, these case could also be explained as processes rather than individual states. Mindless behavior is in many cases caused by a bundle of connected states, unwanted emotions relate to other internal states (which make them unwanted), and suppressed thoughts are suppressed due to other internal states.

Nonetheless, if Schooler means by “state” the “general state of mind”¹⁹ rather than individual states, his examples become more convincing. But this seems inconsistent. Schooler takes inspiration from Dennett, who is interested in beliefs subjects have about phenomenal experience of individual states. Schooler switches between talk of phenomenal experience of individual states, and talk about the stream of consciousness the subject experiences. This is evident in the way he introduces the core distinction, namely in terms of the phenomenal experience of a state. At other times he talks about states of which I am aware, and sometimes about “what is going on in one’s mind”, which I take to refer to the stream of consciousness, or more precisely the sequence or combination of contents of individual states, rather than a classification of the experience of just one state. So the pressing question is really: what kind of reflection is “taking stock” exactly? How should we characterize what we do when we “take stock” and reach meta-awareness? In the following section I present more detailed suggestions for a taxonomy of different kinds of reflection. For now let me just say that one possible view would be that the content of these states (or the states) are accessed by other states, and maybe (unconsciously) evaluated. In that case, we should talk about complex processes rather than re-accessed individual states. Such a view would also be compatible with certain higher-order theories of consciousness.

Later in the paper, Schooler discusses examples of misrepresentation, in his terminology “translational dissociations”: emotions, or cases in which it is less controversial whether a phenomenal character is involved than in case of thoughts. He gives two examples of such misrep-

¹⁹ As formulations such as “take stock what’s going in their own minds” (Schooler this collection, p. 8) suggest.

resentations: emotions of anxiety, which are not reported, and reported disgust for homosexuality. In his first example we find a correlation with the inconsistent behavioral measures of heart rate and galvanic skin response, as indicators of existing unreported anxiety. In his second example we have a correlation with penile tumescence (an erection). In both cases we know the bodily aspects of the emotion well (or the caused bodily changes associated with the feeling on an emotion), and thus, so the argument goes, have evidence for the occurrence of the emotion. But in both cases there is also a discrepancy between the subject's reports (assuming the subject is honest) and its potential reportability. Schooler interprets the behavioral facts as indication of the real emotion the subjects experiences, but in the first case fails to acknowledge, and in the second misinterprets.

I am not so sure. First, the theory of emotion one feels committed to certainly plays a central role. Schooler seems to presuppose that unconscious emotions are not possible. Furthermore, it seems to me that both cases are open to a different interpretation, in fact the same interpretation I suggested for mind-wandering. Both unreported (or unreportable?), emotions of anxiety and reported disgust for homosexuality are complex cases. It might very well be that we do not have an individual content of a state that differs, but we rather simply struggle with a number of different but conflicting emotions, the reported one simply being in conflict with others. In both cases we have rather complex scenarios. And if one defends a multi-component account of emotions, it might very well be that the components of these emotions differ—it could be an element in a network that realizes the state, instead of the content of an individual state. This might seem like a minor point, but I think it is important. It undermines a central second part of the strategy, namely the empirical support for the theoretical distinction. Schooler needs more than a theoretical distinction (his first claim); he needs to show that this very distinction is helpful for understanding certain aspects of consciousness, mind-wandering, and other cases (his second claim; see p. 3). Otherwise the conclusion he draws, the new

methodological approach to studying consciousness, would not follow or would lose its plausibility. So undermining Schooler's second claim by showing that in the case of his examples related to emotions (as well as in case of mind-wandering) this evidence is not as clear as one might think, results in a problem for his view.

But there is another important issue here. The empirical evidence seems to be relevant to the stream of consciousness rather than to the experience versus meta-awareness of individual intentional states. The formulation of the main claims suggests that state consciousness is the issue. However, in other sections Schooler refers to the stream of consciousness (See quote above, p. 8). If this is correct, Schooler's empirical project, or more precisely the evidence he has gathered, is about a central aspect of *creature consciousness*. Philosophers distinguish creature consciousness from mental-state consciousness: the first is about a subject that is conscious (either in general or of something in particular), whereas state-consciousness is about conscious states of a creature that it is conscious. Though Schooler's project (especially claim (1)) is formulated in terms of *state consciousness*, the empirical support targets a different kind of consciousness. This also undermines Schooler's second claim by showing that the meaning of consciousness differs in claims (1) and (2). But, as I pointed out in section 1, the stream of consciousness claim would be compatible with a more functional interpretation of claim (1) as well. There is a way to revise claim (1) in a way that avoids this problem.

Using mostly the empirical evidence of mind-wandering, Schooler then suggest a bundle of criteria we might use for the third-person evaluation of what is actually going on in somebody's mind; in my analysis of his main argument this is the third step (see p. 3). These behavioral criteria include behavioral measures (eye-movements, reading comprehension, sustained attention to response) and neurocognitive criteria (ERP, fMRI, behavioral, neuroscientific, fMRI and others). His list is in the spirit of a cognitive account, and similar to others (Seth et al. 2005; Seth et al. 2008). For protagonists of non-cognitive accounts there seems

to be room for attack. But, as I have mentioned, this is not my project (see p. 2). In this commentary, I prefer to focus on conceptual issues *within* cognitive accounts, rather than the debate between cognitive vs. noncognitive accounts (See p. 2). As long as one commits to such a cognitive account, Schooler's list of criteria turns out to be very useful for our evaluation of the meta-components we need for a fined-grained understanding of reflection and re-representation. And this is the case independently of the worries I presented regarding his first two claims. However, I think there is a problem looming: Schooler is challenging both the reliability of first-person reports and the view that conscious states are accessible states. With a position that is in such sharp tension with our commonsense understanding, he needs to motivate this radical move: he needs to provide an answer to *why* we have this deep pre-theoretic entrenchment of the first-person accessibility of our own conscious states (Cohen & Dennett 2011).

5 A new taxonomy of different kinds of reflection

It's time for a positive proposal. I claimed that I would introduce suggestions for the building blocks of a new taxonomy of different kinds of reflections. As I argued, we need to further specify the kind of reflections required for Schooler's "pure experience" and for his meta-awareness, and to get a better grasp on what is meant by "taking stock" and "re-representation". I also argued that the difference between consciousness and meta-awareness should not be understood as a dichotomy. Rather, we should understand reflection itself as a hierarchical and multidimensional process. So, what exactly is the "taking stock" required for meta-awareness? According to Schooler, meta-awareness requires an explicit representation of the current contents of thought (2011, p. 321). But at least two of the terms involved in this characterization are used in several and distinct meanings: knowledge, and explicit representation.²⁰ Explicit rep-

resentation might be interpreted as being itself conscious, or as having symbolic or conceptual content. The notion of knowledge is also problematic, simply because knowledge is a factive verb. It implies that we cannot be wrong.²¹ As a result, Schooler built the impossibility of misrepresentation into his definition of meta-awareness. This might be consistent with his claim that the first-level perspective inhabits its own ontological realm. But it also creates a problem, because any view that understands introspection or reflection as an inner perception or re-representation automatically has to allow that this process can go wrong. In other words, it has to allow for misperception/misrepresentation. Moreover, Schooler himself want to allow for a certain kind of misrepresentation, in his terminology translational dissociations—cases in which at the personal level we misinterpret what we experience.

In my discussion of the distinction I claimed that we are able to rule out two kinds of reflections that are not helpful. First, categorization under concepts is one kind of meta-cognitive reflection that is itself unconscious, but necessary for conscious experience. Second, being able to distinguish between self and world is another dimension of reflection, at the highest level, that seems necessary for any conscious experience. Neither of these can be the kind of reflection that distinguishes Schooler's experience from meta-awareness. In addition I claimed that both the self-other distinction and self-awareness can happen in a number of ways. Different kinds of meta-cognition, the general ability to monitor and control one's own cognition, and the ability to attribute mental states to oneself and others, can as a result be further specified and characterized along those dimensions.

But again, what is the kind of reflection or "taking stock" required for meta-awareness? At the end of the last section I suggested that the kind of reflexion involved in "taking stock" could be characterized as a case in which the content of these states (or the states) are accessed by other states, and maybe (unconsciously) evaluated. So at issue are complex pro-

²⁰ For a more detailed discussion of the same issue see Metzinger (2013, p. 11).

²¹ Otherwise we would have a false belief, not knowledge.

cesses rather than re-accessed individual states. Such a view is compatible with certain higher-order theories of consciousness. And this would allow that misrepresentation is in fact possible. However, not only do authors like Rosenthal build several meta-representational levels into their theories, the content of the higher level thought contains an element of self, a reference to “oneself”. Self-awareness is built in the analysis, not just any kind of reflection, access, or re-representation. This interpretation uses a certain reading of creature consciousness. It requires that an organism is not only aware but also self-aware. This is a notion of creature consciousness that at first seems to be in tension with Schooler’s main distinction. However, as I argued, this is not necessarily the case. Self-awareness itself comes in degrees and varies along multiple dimensions. Creature consciousness in mind-wandering can than be understood as an intentional relation between the organism and some object or item of which it is aware, in our case a train of thoughts (and/or the sensory states associated with it). This is where the contrast between content theories and functional theories comes into play. As I have argued, pure content or representationalist theories, which claim that conscious states have their mental properties due to their representational properties, are not a good strategic partner for Schooler. In contrast, a certain class of functional accounts, especially higher-order theories, turn out to be a nice fit for his account. These accounts analyze consciousness as a certain form of self-awareness. As a result, we can grant that for the experience of mind-wandering without meta-awareness there is some self-awareness required, and for meta-awareness another more demanding kind of self-awareness is necessary. Rosenthal’s higher-order account would give Schooler this kind of distinction: meta-awareness would include a third-order state, in his terminology a re-re-representation, whereas the experience of mind-wandering would involve only a second-order state, a re-representation (see p. 10).

The literature on phenomenology offers more helpful distinctions of how we can further evaluate these different dimensions. Most of

these distinctions are orthogonal. Several authors claim that “what is likeness” comes in different forms. For example, Carruthers distinguishes the “what it’s likeness” of the world (or *worldly* subjectivity—what the *world* is like for the subject—from *experiential* subjectivity—what the *subject’s experience is like for the subject*; Carruthers 1998, 2000). Rosenthal uses a similar distinction. He distinguishes thin from thick phenomenality, whereby thin phenomenality is the occurrence of a certain qualitative character. Thick phenomenality is richer: “[t]hick phenomenality is just thin phenomenality together with there being something it’s like *for one* to have that thin phenomenality” (Rosenthal 2002, p. 657, emphasis added). So thick phenomenality includes a certain kind of reflexion or extra level; it includes an awareness of a richer kind. For Rosenthal this is identical with the existence of an appropriate higher-order representation. But it is a specific kind of meta-cognitive process, one that contains a representation of “oneself”, or in other terminology, a selfmodel (Metzinger 2003). But the self-model itself, our understanding of ourselves and of the difference between oneself and others, might itself come in degrees and on different levels. So the issue is not meta-cognition or reflexion in general, but different levels and involvements of self-awareness.

Instead of focusing on the differing phenomenology, one might also try to specify the notion of access in further detail (Kouider et al. 2010), a suggestion that I think helps us to better understand what is meant by states referring to other states or accessing them. In the workspace model discussion a simple distinction between cognitive access and cognitive accessibility is introduced to defend the possibility of the abovementioned third category, unaccessed but conscious states. Instead of just access vs. accessibility, I suggest that we distinguish different kinds of access (see p. 11). Rather than assuming a rich phenomenology and differing forms of consciousness, one could also propose that awareness itself might come in degrees and that something like partial awareness might exist (Kouider et al. 2010). Instead of distinguishing dissociable forms of consciousness or differ-

ent kinds of personal level phenomenal character like the above accounts, Kouider et al. (2010) use sub-personal descriptions explaining what awareness might be. More exactly, dissociable levels of access are distinguished and differentiated by a hierarchy of representational levels. In case of partial awareness, we have informational access at *some but not all representational levels*. The crucial idea is that information at other levels can remain inaccessible. Or, in some situations, information at these levels could be accessed, but plausible content is filled, which than potentially results in misrepresentation.

I prefer this line of thinking, and I believe it gives us an improved understanding of the sub-personal processes involved in the different levels of reflection and “taking stock” we want to characterize. This framework is very suitable for a revised understanding of Schooler’s main distinction. However, Kouider et al. (2010) postulate partial access as an alternative explanation for conscious visual perception,²² not for internal cases like mind-wandering. But I think the analyses might be useful for our purposes as well. According to this framework, accessible contents at each level of representation are seen as resulting from the integration of signals with contextual prior information, processes that are also influenced by other internal factors (for example attentional factors or vigilance); this integration is further assumed to be modulated by the degree of confidence of the subject. The result is a more fine-grained perspective on conscious experience; instead of simply conscious or unconscious, we can talk about different dimensions of experience. And this is done at the sub-personal level by a specification of access. This also avoids another problem. As I pointed out, Schooler’s account seems very close in spirit to modern versions of workspace accounts. However, these accounts typically assume all-or-nothing mechanisms for access. This is no prob-

lem for Schooler, who proposes his core distinction as a dichotomy. However, it is a potential problem for the revised view I suggest. But I think this can in fact be an advantage. We can indeed grant that representations *within* each level might be accessed in an all-or-nothing manner (as is assumed in workspace models), but none the less insist that the full set of all the representations associated with this process do not have to be conscious.

Different terminologies aside, I think this fits nicely with the spirit of Schooler’s general distinction, and his distinction between experience and meta-awareness. I admit that these are just first steps towards a better conceptual understanding. But interpreted this way, there is not just conscious experience of mind-wandering versus meta-awareness. The situation is more complex. Reflection comes in many forms and involves representations at many levels, as well as access at all these levels of representation. In addition, whether, and to what degree, self-awareness and a self-model is involved makes a difference as well.

6 Perceptual perspective shifting. The Analogy and the mind-body problem

Let me take stock. I have been through the claims made in Schooler’s main underlying argument (see p. 3) So far, I have discussed claim (1), the initial conceptual distinction between experiential consciousness and meta-awareness, a distinction Schooler sets up as a dichotomy. I then discussed his second claim that there is empirical evidence that this conceptual distinction is something we find in cognitive abilities in cases of dissociations, especially in temporal dissociations like mind-wandering and in emotional transitional dissociations. I then argued that his empirical criteria developed in claim (3) are useful for cognitive accounts in general, independently of the worries one might have about claims (1) and (2). Finally I suggested a finer-grained conceptual distinction between different levels of awareness, different kinds of reflection, and “taking stock”. I will turn now to the last issue we shall examine, which is Schooler’s last and main claim (see p. 3):

²² Like most authors, they focus for the most part on the discussion of conscious perception, and especially Sperling (Block this collection; Fink this collection) and Stroop’s paradigms (see Mroczko-Wąsowicz this collection) and what we can learn from them for consciousness. For a more detailed discussion of the pros and cons or an understanding of consciousness as graded within conscious perception see the debate between Cleeremans (2008), Sergent & Dehaene (2004), Seth et al. (2008), and Overgaard et al. (2006).

Schooler claims that this can be used for a new theoretical and ontological framework for studying consciousness, and this is the declared goal of the target paper. He claims that his perceptual perspective-shifting analogy, together with insights from the sections before, gives us a new ontological perspective on the mind-body problem, not just a new methodological strategy. I found this section of the paper surprising. In my opinion, it is relatively independent of the main project he undertakes. Schooler starts by describing the main thought experiments in the philosophical literature used to challenge reductive physicalism.²³ He concludes that the main problem with the reductive positions is that it needs to “reject” those aspects of first person experience “that are not readily handled by a third-person account” (Schooler this collection, p. 25).

I am not convinced that this is correct. It seems a viable alternative solution to me to just subscribe to the traditional reply, and point to some kind of epistemic gap between the third-person approach and the first-person approach instead of an ontological one. One can admit that there is a gap, but it is an explanatory gap between physical processes and conscious experience. One could even state that the gap may be uncloseable in principle, but that consciousness is nonetheless physical (Levine 1983). That is, there is an epistemological gap, but no ontological gap. That we intuitively see a gap might be true; it does not follow that there actually *is* a gap in what exists. All one can conclude is that, epistemologically, there is gap. In addition, our intuitions might simply be wrong: we might be “innate dualists” and that this is the reason why to so frequently slip back in dualist talk (despite knowing better; Papineau 2011). That is the real reason why commonsense intuition pumping thought experiments work so

well. According to this view, the feeling that some part of reality is “left out”, i.e., the “explanatory gap”, arises only because we simply cannot stop ourselves thinking about the mind-brain relation in a dualist way, though this is actually the wrong thing to do. One can be a reductive physicalist without having to reject the phenomenon of conscious experience, despite the fact that we cannot (yet) reduce it or have proper explanations available as to why we experience certain phenomena the way we do. We can experience a gap, have the intuition that something is “left out”, and nonetheless that very intuition might very well be wrong. I simply do not see the need for Schooler’s solution, the postulation of a new realm, that gives rise to both the physical and subjective reality.

I am also not sure about the meaning of the perceptual perspective-shifting analogy itself. Because it rests on a purely metaphorical use of “perspective”, the analogy does not go through. Perceptual perspective-shifting happens at a personal level, moreover, shifting *experience* at the personal level. The supposedly analogous case occurs at the level of theories or accounts, which emphasize either the first- or third-person perspective. But individual experiences differ in principle from the focus certain theories have. Schooler suggests that the resolution of the conflicting perspectives lies in a meta-perspective that acknowledges the existence and irreducibility of both, even though both are somehow equally valid, such that the solution to this tension is a new realm, a meta-perspective which gives us a “higher-order outlook” (Schooler this collection, p. 26). However, Schooler agrees “that [i]t is easier to recognize the need for a meta-perspective than to identify precisely what such a view might be” (Schooler this collection, p. 26). He admits the character of the introduced meta-perspective is “speculative and highly underspecified” (Schooler this collection, p. 28) but thinks that it has intuitive appeal. He also concedes that this is the most speculative part of the paper. I must admit that I struggle with the concept. I fail to see the intuitive appeal. Mostly because it eludes my understanding what the proposed meta-perspective might be and how it is help-

²³ However, this section of the target paper goes beyond the discussion of the well-known traditional arguments from the philosophical debate, including the explanatory gap argument, the Mary argument, and others. Schooler adds a section on the phenomenon of time experience and reductive accounts that explain time. I found the last example very inspiring, because in contrast to the other arguments it is not just based on thought experiments. However, the implications of this for my purposes here do not matter; they are used as a intuition pump to appeal to the necessity of a meta-level, so I cannot cover this aspect in this commentary.

ful despite acknowledging our commonsense intuition, not at an epistemological but an ontological level. As a result, I do not find it explanatory. Moreover, it does not follow from the analogy. For an argument by analogy one needs properties shared by both parts of the analogy. Even if we admit that in perceptual perspective-shifting both personal-level interpretations of an ambiguous figure are equally valid, it does not seem to follow that the first-person perspective and the third-person perspective in *strategies to study consciousness* require a meta-perspective not identical with either of these perspectives. Both the cases seem to have only one thing in common, “perspective shifting”. But “perspective” is used purely metaphorical in the second case. Moreover, bridging the first- and third-person perspectives seems to be an epistemic challenge. But from an epistemic observation or claim an ontological claim does not follow. Even if we admit an epistemic gap and agree that we cannot help but see an explanatory gap in all these cases, the postulation of an independent higher-order meta-level, an ontological claim, is not well supported. In addition, both of these issues, the ontological as well as the epistemological claim, differ from the methodological approach defended by Schooler. To summarize, in my opinion this section, and the preferences regarding solutions of the mind-body problem, are conceptually relatively independent from the main project, which I take to be the development of a useful strategy to study consciousness and mind-wandering. Schooler’s strategy might be helpfully independently of whether one is a reductive or non-reductive physicalist. I think that such a methodological reading of his approach strengthens the project, because it disassociates it from a completely different issue.

7 Conclusion

Having noted the initial plausibility of the general outline of Schooler’s account, I pointed out some problems and expressed some general reservations about its scope. First, I argued that the postulation of a third kind of

conscious but not accessed or reflected state is not justified. As a result, the account is too narrow, because one of the underlying general assumptions is not justified. This assumption causes a number of problems and a few misunderstandings. However, the assumption seems conceptually independent of the main project, which is to allow us to bridge the gap between first- and third-person criteria for consciousness. I suggested that the main distinction is underspecified and needs further clarifications of the elements involved: access, reportability, and levels of awareness.

Second, although it is tempting to attribute a first-order account to Schooler, a more convincing alliance would actually be certain functional accounts, especially higher-order accounts and global workspace accounts. And I argued that we should replace the introduced dichotomy by a finer-grained distinction of different kinds of meta-cognitive processes and meta-reflections in several dimensions.

In discussing support for the underlying conceptual framework, I then argued that the evidence offered is actually about complex cases. As exciting as the empirical results are, they seem not to be about individual states, but rather about the connection between many states or even the stream of consciousness. The project is about creature consciousness, not state consciousness—though the initial distinction suggests otherwise. This is the first result of my commentary.

I would suggest giving up the idea that the account offers a new meta-perspective, which for Schooler is a preferable alternative to reductive physicalist accounts. I do not think there is a need for this ontological element in his account, and it does not seem to fit with the rest of the methodological project. In addition, the claim seems independent of the rest of the project and there are reductive accounts available that fit very nicely with his project. This is the second result.

In essence I suggested a few conclusions and recommendations, mostly based on conceptual considerations, which clarify and strengthen the main project, with which I sympathize.

1. We keep many main insights of the paper:
 - a) The account is still be a cognitive account, and we allow that cognitive factors help to get a grasp on consciousness; the project is still to bridge the gap between the first- and third-person perspective.
 - b) We also keep the insight that further processing and certain kinds of further processes might either change the state itself and/or the state's content. But we acknowledge that we need to consider the embeddedness of the state to determine the experience. In other words, we focus on processes and phenomena, instead of individual states. This allows Schooler to associate his project with either a hybrid account or a version of a functional account, more specifically a workspace account or higher-order account. Which in turn helps to specify the dimensions of meta-processing in more detail and get a better grasp of the necessary conceptual clarifications. Nonetheless, we still see meta-awareness and consciousness as distinct phenomena. I take this to be the driving idea in his initial distinction.
 - c) The proposed list of potential criteria is still extremely useful, since it helps to determine these very reflective dimensions and factors, which determine both experience and the activities of the mind. For example, the behavioral criteria²⁴ will be caused by these very meta-processes, which we try to identify in more detail.
 - d) Finally, we keep the insight that factors accessible through the third-person perspective can give us insight into what is going on in the mind, as well as in conscious processes.
2. The remaining task, then, is to specify the aspects and dimensions that are relevant, and the kinds of meta-processes, access, or reflection in question. I suggested building blocks for an improved taxonomy of different kinds of reflections and "taking stock". I suggested that awareness itself might come in degrees and at different levels of representation. By

distinguishing dissociable levels of access differentiated at hierarchical representational levels, we allow for partial awareness. In effect, this allows for a fine-grained perspective on conscious experience. Instead of just unconscious, conscious, and a meta-reflective level of awareness, we have different dimensions of experience. And this is done at the sub-personal level by a specification of the term "access". But we should restrain from simply postulating a third category, namely a state that is unaccessed (or un-accessible) but conscious, thereby avoiding the problems associated with the postulation of this third category. The resulting finer-grained taxonomy allows an improved understanding of how exactly meta-awareness and conscious experience differ. Of course there is a price to pay if we accept this change of focus. While we can still claim that the criteria give an insight into what is going on *in the mind*, "the mind" includes unconscious states, conscious states, and several levels of re-representational processes.

There are a number of advantages of a view like this. First, it is not in conflict with some of the most promising candidates for philosophical theories of consciousness. Moreover, one can still account for the similarity of an unconscious state and its conscious counterpart. And third, one can keep the initial idea behind Schooler's distinction between the experienced state and a meta-reflective level of awareness of "knowing that one is in this state", but would substitute it with a finer-grained conceptual framework of multiple differences among several dimensions.

²⁴ For example in the the discussion of emotions p. 13.

References

- Baars, B. J. (1988). *A cognitive theory of consciousness*. Cambridge, UK: Cambridge University Press.
- (1996). *In the theater of consciousness: The workspace of the mind*. New York, NY: Oxford University Press.
- Bayne, T. & Montague, M. (Eds.) (2011). *Cognitive phenomenology*. New York, NY: Oxford University Press.
- Block, N. (2005). Bodily sensations as an obstacle for representationalism. In M. Aydede (Ed.) *Pain. New essays on its nature and the methodology of its study* (pp. 137-142). Cambridge, MA: MIT Press.
- (2011). Perceptual consciousness overflows cognitive access. *Trends in Cognitive Sciences*, 15 (12), 567-575. [10.1016/j.tics.2011.11.001](https://doi.org/10.1016/j.tics.2011.11.001)
- (2015). The puzzle of perceptual precision. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Boyd, R. (1999). Kinds, complexity and multiple realization. *Philosophical Studies*, 95 (1/2), 67-98.
- Braddon-Mitchell, D. & Jackson, F. (2007). *Philosophy of mind and cognition*. Oxford, UK: Blackwell.
- Byrne, A. (2001). Intentionalism defended. *Philosophical Review*, 110 (2), 199-240. [10.1215/00318108-110-2-199](https://doi.org/10.1215/00318108-110-2-199)
- Carruthers, P. (1998). Natural theories of consciousness. *European Journal of Philosophy*, 6 (2), 203-222. [10.1111/1468-0378.00058](https://doi.org/10.1111/1468-0378.00058)
- (2000). *Phenomenal consciousness*. Cambridge, UK: Cambridge University Press.
- (2005). Conscious experience versus conscious thought. In P. Carruthers (Ed.) *Consciousness: Essays from a higher-order perspective*. Oxford, UK: Oxford University Press.
- Carruthers, P. & Veillet, B. (2011). The case against cognitive phenomenology. In T. Bayne & M. Montague (Eds.) *Cognitive phenomenology* (pp. 35-56). New York, NY: Oxford University Press.
- Cleeremans, A. (2008). Consciousness: The radical plasticity thesis. *Frontiers in Psychology*, 168, 19-33. [10.1016/S0079-6123\(07\)68003-0](https://doi.org/10.1016/S0079-6123(07)68003-0)
- Cohen, M. A. & Dennett, D. C. (2011). Consciousness cannot be separated from function. *Trends in Cognitive Sciences*, 15 (8), 358-364. [10.1016/j.tics.2011.06.008](https://doi.org/10.1016/j.tics.2011.06.008)
- (2012). Response to Fahrenfort and Lamme: Defining reportability, accessibility and sufficiency in conscious awareness. *Trends in Cognitive Sciences*, 16 (3), 139-140. [10.1016/j.tics.2012.01.002](https://doi.org/10.1016/j.tics.2012.01.002)
- Dehaene, S., Changeux, J.-P., Nacchache, L., Sackur, J. & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: A testable taxonomy. *Trends in Cognitive Sciences*, 10 (5), 204-211. [10.1016/j.tics.2006.03.007](https://doi.org/10.1016/j.tics.2006.03.007)
- Dretske, F. (1993). Conscious experience. *Mind*, 102 (406), 263-283.
- Fahrenfort, J. J. & Lamme, V. A. F. (2012). A true science of consciousness explains phenomenology. Comment on Cohen and Dennett. *Trends in Cognitive Sciences*, 16 (3), 138-139. [10.1016/j.tics.2012.01.004](https://doi.org/10.1016/j.tics.2012.01.004)
- Fink, S. B. (2015). Phenomenal precision and its possible pitfalls: A commentary on Ned Block. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Gennaro, R. J. (1996). *Consciousness and self-consciousness: A defense of the higher-order thought theory of consciousness*. Amsterdam, NL: John Benjamins.
- Grush, R., Jaswal, L., Knoepfler, J. & Brovold, A. (2015). Visual Adaptation to a Remapped Spectrum. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1-16). Frankfurt a. M., GER: MIND Group.
- (2015). Visual adaptation to a remapped spectrum: Lessons for enactive theories of color perception and constancy, the effect of color on aesthetic judgments, and the memory color effect. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Jackendoff, R. S. (1987). *Consciousness and computational mind*. Cambridge, MA: MIT Press.
- Kouider, S., de Gardelle, V., Sackur, J. & Dupoux, E. (2010). How rich is consciousness? The partial awareness hypothesis. *Trends in Cognitive Sciences*, 14 (7), 301-307. [10.1016/j.tics.2010.04.006](https://doi.org/10.1016/j.tics.2010.04.006)
- Kouider, S., Sackur, J. & de Gardelle, V. (2012). Do we still need phenomenal consciousness? Comment on Block. *Trends in Cognitive Sciences*, 16 (3), 140-141. [10.1016/j.tics.2012.01.003](https://doi.org/10.1016/j.tics.2012.01.003)
- Kriegel, U. (2009). *Subjective consciousness: A self-representational theory*. Oxford, UK: Oxford University Press.
- Lamme, V. (2003). Why visual attention and awareness are different. *Trends in Cognitive Sciences*, 7 (1), 12-18.
- Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, 64, 354-361.
- Lycan, W. (1996). *Consciousness and experience*. Cambridge, MA: MIT Press.

- Metzinger, T. (1995). Faster than thought: Holism, homogeneity and temporal coding. In T. Metzinger (Ed.) (pp. 425-461). Imprint Academics.
- (2003). *The self-model theory of subjectivity*. Cambridge, MA: MIT Press.
- (2013). The myth of cognitive agency: subpersonal thinking as a cyclically recurring loss of mental autonomy. *Frontiers in Psychology*, 4 (931), 1-19. [10.3389/fpsyg.2013.00931](https://doi.org/10.3389/fpsyg.2013.00931)
- Mroczko-Wąsowicz, A. (2015). What can sensorimotor enactivism learn from studies on phenomenal adaptation in atypical perceptual conditions? A commentary on Rick Grush and colleagues. In T. Metzinger & J. M. Wiandt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Nelkin, N. (1989). Propositional attitudes and consciousness. *Philosophy and Phenomenological Research*, 49 (3), 413-430.
- Northoff, G. (2014). *Unlocking the brain. Vol II: Consciousness*. New York, NY: Oxford University Press.
- Overgaard, M., Rote, J., Mouridsen, K. & Ramsøy, T. Z. (2006). Is conscious perception gradual or dichotomous? A comparison of report methodologies during a visual task. *Conscious Cognition*, 15 (4), 700-708.
- Overgaard, M. & Grünbaum, T. (2011). Cognitive and non-cognitive conceptions of consciousness. *Trends in Cognitive Sciences*, 16 (3), 137-137. [10.1016/j.tics.2011.12.006](https://doi.org/10.1016/j.tics.2011.12.006)
- Papineau, D. (2002). *Thinking about consciousness*. Oxford, UK: Oxford University Press.
- (2011). What exactly is the explanatory gap? *Philosophia*, 39 (1), 5-19. [10.1007/s11406-010-9273-6](https://doi.org/10.1007/s11406-010-9273-6)
- Pitt, D. (2004). The phenomenology of cognition. Or what is it like to think that P? *Philosophy and Phenomenological Research*, 69 (1), 1-36. [10.1111/j.1933-1592.2004.tb00382.x](https://doi.org/10.1111/j.1933-1592.2004.tb00382.x)
- Prinz, J. J. (2004). The fractionation of introspection. *Journal of Consciousness Studies*, 11, 40-57.
- (2012). *The conscious brain*. New York, NY: Oxford University Press.
- Robinson, W. S. (2005). Thoughts without distinctive non-imagistic phenomenology. *Philosophy and Phenomenological Research*, 70 (3), 534-561.
- Rosenthal, D. (1986). Two concepts of consciousness. *Philosophical Studies*, 49, 329-359.
- (2002). How many kinds of consciousness? *Consciousness and Cognition*, 11 (4), 653-665. [10.1016/S1053-8100\(02\)00017-X](https://doi.org/10.1016/S1053-8100(02)00017-X)
- (2005). *Consciousness and mind*. Oxford, UK: Oxford University Press.
- (2012). Higher-order awareness, misrepresentation, and function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367 (1594), 1424-1438. [10.1098/rstb.2011.0353](https://doi.org/10.1098/rstb.2011.0353)
- Schooler, J. W. (2002). Re-representing consciousness: Dissociations between experience and meta-consciousness. *Trends in Cognitive Sciences*, 6 (8), 339-344. [10.1016/S1364-6613\(02\)01949-6](https://doi.org/10.1016/S1364-6613(02)01949-6)
- (2013). The costs and benefits of mind-wandering: A review. *Canadian Journal of Experimental Psychology*, 67 (1), 11-18. [10.1037/a0031569](https://doi.org/10.1037/a0031569)
- (2015). Bridging the objective/subjective divide: Towards a meta-perspective of science and experience. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Schooler, J. W., Smallwood, J., Christoff, K., Handy, T. C., Reichle, E. D. & Sayette, M. A. (2011). Meta-awareness, perceptual decoupling and the wandering mind. *Trends in Cognitive Sciences*, 7 (15), 319-326.
- Sergent, C. & Dehaene, S. (2004). Is consciousness a gradual phenomenon? Evidence for an all-or-none bifurcation during the attentional blink. *Psychological Science*, 15 (11), 720-728. [10.1111/j.0956-7976.2004.00748.x](https://doi.org/10.1111/j.0956-7976.2004.00748.x)
- Seth, A. K., Baars, B. J. & Edelman, D. B. (2005). Criteria for consciousness in humans and other mammals. *Consciousness and Cognition*, 14 (1), 119-139. [10.1016/j.concog.2004.08.006](https://doi.org/10.1016/j.concog.2004.08.006)
- Seth, A. K., Dienes, Z., Cleeremans, A., Overgaard, M. & Pessoa, L. (2008). Measuring consciousness: Relating behavioural and neurophysiological approaches. *Trends in Cognitive Sciences*, 12 (8), 314-321. [10.1016/j.tics.2008.04.008](https://doi.org/10.1016/j.tics.2008.04.008)
- Siegel, S. (2005). Which properties are represented in perception? In T. Gendler & J. Hawthorne (Eds.) *Perceptual Experience* (pp. 481-503). Oxford, UK: Oxford University Press.
- Siewert, C. (1998). *The significance of consciousness*. Princeton, NJ: Princeton University Press.
- Tye, M. (1995). *Ten problems of consciousness*. Cambridge, MA: MIT Press.
- Tye, M. & Wright, B. (2011). Is there a phenomenology of thought? In T. Bayne & M. Montague (Eds.) *Cognitive Phenomenology* (pp. 326-344). New York, NY: Oxford University Press.
- Van Gulick, R. (2000). Inward and upward: Reflection, introspection, and self-awareness. *Philosophical Topics*, 28 (2), 275-305.

Stepping Back and Adding Perspective

A Reply to Verena Gottschling

Jonathan Schooler

In this reply, I circumvent (some might say dodge) a number of Gottschling's fine-grained comments by stepping back and reviewing the key points of the three major sections of my target paper in light of her more general concerns. I first consider Gottschling's primary criticism of the first section of my paper, namely that insights that might emerge from considering the perspective shifting associated with reversible images do not apply in the context of differences between first and third-person perspectives. Although I concede there are differences in the meaning of "perspective" in conceptual and perceptual domains, I argue that the common element of a reliance on a frame-of-reference is sufficient to make the analogy helpful. I contend that a necessary element in overcoming the limitations of particular perspectives in both conceptual and perceptual domains is attempting to consider alternative vantages. This approach is then used to justify the tack of the next two sections: considering first-person experience from the vantage of third-person science and considering third-person science from the vantage of first-person experience. I note that Gottschling is largely sympathetic to the broad goals of the second section of my paper, and observe that her major concern with the construct of experiential consciousness emerges from her burdening it with unwarranted assumptions. I use her constructive suggestion for the need for further development of the notion of meta-awareness as a springboard for introducing a previously overlooked element (experiential monitoring) that may be useful for explaining how people can knowingly monitor performance without explicit verbal re-representation. Finally, I consider Gottschling's view that the third section fails to add to the value of the paper. Although I acknowledge that the arguments in the second section stand independently, I argue that discussion of how science can inform experience gains greater balance by also considering how experience informs science. I close by challenging the view that knowledge gained from science necessarily trumps that gained by experience, and conclude that it remains a worthy goal to seek a meta-perspective that accommodates both first- and third-person perspectives without reducing one to the other.

Keywords

Consciousness | Explanatory gap | Frame-of-reference | Heterophenomenology | Meta-awareness | Meta-perspective | Mind-wandering | Mind/body problem | Mindfulness | Monitoring | Neurophenomenology | Neutral monism | Panpsychism | Phenomenology | Time

1 Introduction

Reviewing a commentary on one's work, even one as thoughtful as that provided by [Gottschling \(this collection\)](#), is much like viewing a close-up picture of one's face on a large high-definition screen; every blemish seems patently visible and appears to overshadow even the most genuine of expressions. The temptation is to pull out one's

metaphoric Photoshop and doctor up every imperfection. There is another option, however, and that is to step back and consider whether from a broader perspective the blemishes are really as disfiguring as they might initially appear.

Inspired by this analogy, I will not attempt to rebut all of Gottschling's consistently

Author

[Jonathan Schooler](#)

jonathan.schooler@psych.ucsb.edu
University of California
Santa Barbara, CA, U.S.A.

Commentator

[Verena Gottschling](#)

vgott@yorku.ca
York University
Toronto, ON, Canada

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu
Monash University
Melbourne, Australia

incisive remarks about my paper. Rather I will use this essay as an opportunity to step back and review the broad strokes of my arguments in light of Gottschling's more general concerns. In so doing, I hope to demonstrate that while Gottschling offers a number of insightful suggestions for clarification and elaboration, the general logic of my arguments remain largely intact. Nevertheless, Gottschling's critique offers an excellent opportunity to clarify some points that may have been lost in the expanse of my initial paper.

2 Reflections on section 2¹: Applying perspective shifts to conceptualizing human experience from the first- versus third-person perspective

My paper opens with the contention that seemingly opposing arguments can often reflect alternative vantages of a larger meta-perspective from which both views can be understood. I illustrate this point using the example of reversible images that can be seen as corresponding to two entirely different objects depending on one's perspective. I argue that when one recognizes that both vantages are true from their particular perspective, one gains an understanding of the larger context (i.e., a meta-perspective). Although most of my examples are perceptual illustrations, I suggest that there is a close correspondence between the processes involved in perspective taking in perceptual and conceptual domains, and that an appreciation of meta-perspectives in the perceptual domain may help the formulation of meta-perspectives in the conceptual domain. In the spirit of this argument I suggest that the long-standing debate between approaches that emphasize the subjective first-person perspective of experience and those that emphasize the objective third-person perspective of science, may be akin to debating which direction the dancer is rotating in the spinning dancer illusion (see figure 6 in [Schooler this collection](#)). In both cases, it simply depends on your perspective. Taken from the perspective of

the individual, understanding consciousness necessarily invites a reliance on introspection and first-person analysis. Taken from the perspective of conventional third-person science, understanding consciousness necessarily requires objectively observable facts (e.g., behaviors, physiological responses) that can be derived independently of any single individuals' experience

I argue that both of these views have merit, that both researchers and schools of thought have debated (often vehemently) about which of these two vantages is more appropriate, and that part of the heat of this controversy may stem from people's disinclination to switch back and forth between perspectives and thereby gain a larger view that treats neither as decisively superior.

Gottschling rejects the notion that the alternative perspectives afforded by reversible images has relevance to conceptualizing the challenges of reconciling first- and third-person perspectives. Her difficulty with this analogy stems (at least in part) from her view that the meaning of "perspective" in these two contexts does not align. As she puts it: "Because it rests on a purely metaphorical use of 'perspective', the analogy does not go through" ([Gottschling this collection](#), p. 18). To be sure there are significant differences between the meaning of "perspective" in the context of perceptual experience, such as reversible images, and conceptual ideas, such as the difference between first- and third-person approaches to the study of consciousness. However, I argue that there are some deep parallels between the meanings of "perspective" in these two contexts that make the analogy a useful one. I'll begin by considering the broader issue of the parallels between perceptual and conceptual perspectives and then the more specific question of how these parallels might usefully apply to the conceptual distinction between first- and third-person perspectives.

Critically, in both perceptual and conceptual contexts "perspective" is defined by a frame-of-reference that determines how the constituent elements are understood and related to one another, as well as which elements are

¹ The paper begins with a very brief introduction that is numbered section 1. As a result the first major section of the paper is numbered section 2.

taken as central and which as more peripheral. In perceptual contexts, the frame-of-reference is defined in terms of the assignment of spatial arrangements; i.e., what is to the left and the right, what is in the foreground and background etc. In conceptual contexts, the frame-of-reference is defined in terms of the assignment of conceptual arrangements; i.e.; which elements are conceptually closer or further apart, which are more essential and which more peripheral. In both cases, frame-of-reference can have profound effects as evidenced by the reversible image research in perception (Chambers & Reisberg 1992) and research on cognitive framing (Tversky & Kahneman 1981) in cognition. A further striking parallel between perceptual and conceptual perspectives is that they both become easily entrenched. When one watches the spinning dancer (figure 6) it is very difficult to recognize that at any time she can be seen as facing in one of two different directions. In a very similar way, when one works on a conceptual problem it is very easy to interpret it in a particular way that creates a “mental set” that can impede its solution. There is even a common cognitive ability (Schooler & Melcher 1995; see also, Wiseman et al. 2011) for overcoming the mental sets associated with solving conceptual problems (e.g., insight problems) and perceptual problems (e.g., recognizing out-of-focus pictures). In short, perceptual reversible images elegantly illustrate a fundamental aspect of not just perception but of human cognition more generally; namely, that we routinely consider things (be they objects or ideas) within the context of a particular frame-of-reference (be that frame perceptual or conceptual), and we can have a very hard time reconsidering those things from a different perspective.

Even if it is appropriate to draw a parallel between the meaning of “perspective” in perceptual and conceptual contexts, it does not necessarily follow that the analogy can be extended to the particular conceptual problem of distinguishing between the first- and third-person perspective approaches. But I maintain that it is in fact particularly applicable in this context. The essence of the distinction between first- and third-person perspectives has to do with one’s

frame-of-reference. If one considers consciousness from a first-person perspective, one is understanding it in relationship to one’s own personal experience, taking subjectivity as the foreground and objective reality as the background. One is considering consciousness through one’s own experience, and grounding assumptions on what is real and important on the basis of that personal subjective vantage. In contrast, a third-person perspective takes the objective world as the frame-of-reference. Personal experiences that cannot be independently verified are therefore suspect and inferences must be drawn, as they are in all of science, on the basis of people’s measurable behaviors and physiological responses. In my view, it is no accident that these two approaches to thinking about consciousness have historically been described in terms of differences in *perspective* as they self-evidently entail thinking about consciousness from distinctly different frames-of-reference.

In short, I maintain that the notion of distinct conflicting perspectives akin to those associated with perceptual reversible images aptly applies to many conceptual distinctions, but especially apply when it comes to characterizing the objective/subjective divide. The corollary of this claim is the possibility that, like the alternative perspectives of reversible images, the objective/subjective divide may be usefully informed by recognizing that both perspectives represent equally meaningful interpretations that cannot be reduced to one another, but may be better understood from a meta-perspective that acknowledges the larger context in which they are both embedded.

In my view, the importance of the distinct perspectives that emerge from alternative frames-of-reference simply cannot be overstated. In addition to its self-evident effects in the context of perception, frames-of-reference are a powerful determinant of the actions that people take in important real-life situations. For example, doctors’ prescriptions of how to treat an epidemic is profoundly influenced by whether the treatment is framed in terms of lives saved or lives lost even when it corresponds to precisely the same scenario (Tversky & Kahneman 1981). In physics, fundamental breakthroughs

have repeatedly taken place as a function of changes in scientists' frame-of-reference. For example, Newton's laws of gravity emerged when he realized that the same frame-of-reference that applies to forces on the ground equally applies to the motion of the heavens (Westfall 1980). Einstein's special theory of relativity was fostered by his replacement of the notion of an absolute frame-of-reference with a frame-of-reference defined relative to the observer (2001). Given the significance of perspective and frame-of-reference in other contexts it stands to reason that something so salient as whether one is thinking about consciousness from their own perspective or from the objective perspective of science should profoundly impact the questions that they ask and the answers that they reach.

In the case of reversible images, the best way to understand how they can correspond to two so entirely distinct yet self-consistent representations is to practice alternating between vantages. Although at first it is very difficult to see how the spinning dancer alternatively rotates in two different directions, with practice one comes to appreciate the two vantages that the image affords, and thus to understand why her direction changes. The primary goal of my paper is to explore the hypothesis that a deeper understanding of the subjective/objective divide can emerge in a similar fashion. By thoroughly considering each vantage from the perspective of the other, it is hoped that a meta-perspective will emerge that recognizes the logical consistency of each, while not attempting to reduce either one to the other.

Gottschling suggests that my emphasis on "meta-perspective" is an unnecessary strategic move that ultimately detracts from the primary value of my paper. Part of her difficulty with the meta-perspective emphasis may arise from my inadequately situating the second section of my paper in the context of this construct, and the seeming equation of meta-perspective with non-reductionism in the third section. However, the value of considering alternative perspectives in overcoming the limitations that can emerge when one solely considers a single vantage has merit regardless of whether one ascribes to any of the ontological speculations I suggest in the

third section of my paper. Independent of the conclusions that one derives, there seems to be great value in systematically considering subjective experience from the vantage of a third-person perspective, and objective reality from the vantage of a first-person perspective, which are the goals of section 2 and section 3 respectively.

3 Reflections on section 3: Gaining a third-person perspective on people's first-person experience

In the second section of my paper I review research that attempts to inform our understanding of the first-person experience using the third-person perspective of science. This approach takes at its starting point a theoretical distinction between experiential consciousness (corresponding to the contents of on-going experiences) and meta-consciousness (or meta-awareness—the terms are used interchangeably) corresponding to the explicit re-representation of the contents of experiential consciousness. These levels are illustrated by the case of mind-wandering while reading. In this context, experiential consciousness corresponds to the content of the mind-wandering episode and meta-awareness is initially absent but suddenly emerges with the realization that one was mind-wandering rather than attending to the text.

An important implication of the distinction between experiential consciousness and meta-consciousness is that people can have experiences (e.g., mind-wandering) that they either fail to notice explicitly (temporal dissociations) or notice but manage to mischaracterize (translation dissociations). I review a program of research that has fleshed out this distinction in various contexts, with a particular focus on mind-wandering. Using assorted methodologies including the combination of experience sampling measures, self-catching, and behavioral measures, we find evidence that people routinely fail to notice episodes of mind-wandering but are nevertheless accurate at reporting it when they are directly queried.²

² A very recent paper (Seli et al. in press) suggests some variability in the accuracy of mind-wandering reports as assessed by the corres-

Gottschling devotes the bulk of her remarks to discussing efforts to develop a third-person science of first-person experience. In general, she is sympathetic to the approach. However, she raises a variety of concerns and makes a number of useful suggestions. As noted, I will not endeavor to respond to all of her concerns; however, there are several that stand out, and so I will consider them in turn.

Gottschling's primary reservation about the distinction between experiential consciousness and meta-awareness is that she is not persuaded by my characterization of experiential consciousness. Essentially she does not see how it is possible to "distinguish conscious processes which are not accessed from unconscious activity" (Gottschling [this collection](#), p. 11). Although it is true that there are some situations where it may be difficult to distinguish experienced but not meta-aware from unconscious processes (as in the case of potentially unconscious emotions, see [Schooler et al. 2015](#)), often this distinction is quite straightforward. For example, when people are surprised to suddenly realize that they are mind-wandering instead of paying attention to what they reading. In this case, it is evident that they were experiencing the contents of the mind-wandering as they are typically able to report them. It is simply that they had not engaged in the reflective process of noting that they were mind-wandering instead of reading. In short, Gottschling is unpersuaded by a mental state—"conscious processes which are not accessed"—that I never actually postulated. Essentially, she layered onto the construct the notion that experiential consciousness is not accessed, and then criticized it for this reason.

In fact, although I am not committed to the notion that non-conscious higher order thoughts underpin all conscious thoughts ([Rosenthal 1986](#)), I have no problem with Gottschling's attempted revision to my notion of experiential consciousness, namely that it represents a third-order level of consciousness. Indeed I have speculated about this possibility in the

pendence of such reports to behavioral indices of lapses. Nevertheless, people appear to have some access to when their reports are likely to be more vs less accurate as evidenced by a significant correlation between confidence in self-reports and correspondence to the behavioral indices of mind-wandering.

past (see [Schooler et al. 2015](#)). I am therefore entirely comfortable with Gottschling's suggestion that "meta-awareness would include a third-order state, in his terminology a re-re-representation whereas the experience of mind-wandering would involve only a second-order state, a re-representation" ([this collection](#), p. 16). Just so long as the second-order cognition is not *experienced* as a reflection about experience, I have no problems with whatever non-conscious higher-order cognitions may be required to produce it.

Although Gottschling's concerns with the notion of experiential consciousness seem to be largely a product of her reading into my distinction more than was intended, her suggestion that it may be helpful to consider more fine-grained levels of meta-awareness is a worthwhile idea that merits development. As Gottschling observes, there is a need for "an improved taxonomy of different kinds of reflection and 'taking stock' ... awareness itself might come in degrees and at differently levels of representation" ([this collection](#), p 20). Indeed, one feature that has been notably absent from my discussion of meta-consciousness (here and elsewhere) is consideration of the possibility of monitoring processes that may take place at the experiential level, without explicit re-representation at the meta-level. For example, sometimes when people are on-task they may experience a palpable sense of sustained attention without having explicitly to note to themselves that they are on-task. Similarly, when mind-wandering, people sometimes report that they knew they were mind-wandering. This awareness, however, may not necessarily be associated with an explicit acknowledgment of that fact. Rather they maintain a continuous unstated awareness that they are off-task. In short, a further distinction may be needed between a non-propositional "feeling of awareness" that one is doing something ("experiential monitoring") and the verbal/ propositional state of meta-awareness that may occur when people intermittently take stock of their mental state, as when one suddenly thinks to themselves, "Darn! I was mind-wandering again!"

The notion that sometimes people explicitly re-represent their state to themselves (meta-awareness) whereas other times they simply "just

know” they are in that state (experiential monitoring) would also be consistent with alternative mindfulness practices (Thompson 2014). For instance, open-monitoring involves monitoring the content of experience from moment-to-moment without deliberately attending to any particular object (Lutz et al. 2008). Open-monitoring cultivates an aspect of mindfulness described as “observing”, measured with items such as “When I walk, I deliberately notice the sensation of my body moving” (Baer et al. 2006). This seems akin to what I am referring to as experiential monitoring. A somewhat different practice involves labeling one’s experiences as they occur with short tags like “thinking,” “feeling,” or “sensation.” This cultivates an aspect of mindfulness called “describing”, measured with items such as: “My natural tendency is to put my experiences into words.” This process of re-representing experience in words seems akin to meta-awareness.

The distinction between experiential monitoring and meta-awareness might also speak to another of Gottschling’s concerns, namely the question of whether meta-awareness is necessarily all-or-none (as I intimated) or more continuous (as she proposes). Although research would be required to tease out this conjecture, it seems quite plausible to me that experiential monitoring might take place at a continuous level with individuals ranging from either dimly to explicitly aware of what they are doing. In contrast, a more discrete process may occur when individuals suddenly realize that they are engaging in a mental state (e.g., mind-wandering) that they had not previously noticed.

Several other paper concerns that Gottschling raises about my paper, including the possibility of unconscious emotions and how the distinction between experiential consciousness and meta-awareness relates to other distinctions of consciousness (including those of Dehaene et al. 2006; Block 1995 and Rosenthal 1986) are discussed in other locations (e.g., Schooler et al. 2015). While she points out a number of other modest blemishes that I will not address, ultimately the approach for gaining a third-person perspective of first-person experience that I articulated in section 2 of my paper appears logically intact.

4 Reflections on section 4: Toward a meta-perspective for considering the meta-physics of first- versus third-person perspective

Gottschling seems less optimistic about the contribution of the third section of my paper. She dismisses speculations I derive from considering third-person science from the vantage of first-person experience, as a “largely unnecessary strategic move” (Gottschling this collection, p. 1) that “does not seem to fit with the rest of the project” (p. 22). I concur with Gottschling that first person experience can be assessed from the third person perspective of science without also considering objective science from a first-person perspective. In the past I have routinely considered what science has to say about first-person experience without considering the other side of the coin (e.g., Schooler 2002; Schooler et al. 2011; Schooler et al. 2015). Clearly the two sides of the discussion are not logically co-dependent on one another.

I acknowledge that the final section of the paper was not necessary for shoring up any of my arguments in the second section. Nevertheless I maintain that it adds an important balance to the discussion by illustrating the potential value of considering both first- and third-person approaches from the vantage of the alternative perspective. In this concluding section of my paper, I change my frame-of-reference from a third- to a first-person perspective, and consider the current assumptions of science from this vantage. I identify three aspects of existence that I argue are axiomatic from a first-person perspective, including: the existence of experience, the flow of time, and the fact that the present is qualitatively different from the past or the future. I argue that all three of these essential elements are either unexplained by science (i.e., experience) or outright discounted as an illusion of consciousness (i.e., the flow of time, the privileged present). I contend that while many aspects of experience could be illusory, it is hard (indeed impossible for me) to conceive of how experience, the flow of time, or the privileged nature of the present could be among them. On these grounds, I suggest that there may be something missing from the current

account of objective science and speculate that an additional subjective dimension of time might fit the bill. I argue that a subjective dimension of time would provide: 1) a realm of reality for experience to reside, 2) the additional degree of freedom necessary to enable the flow of time in physics' current "block universe", and 3) a way to conceptualize the present. I readily acknowledge that such an account is highly speculative, but I offer it as an example of the type of meta-perspective that I think could emerge by attempting to reconcile the axioms required for both objective and subjective frames-of-reference.

Gottschling's assessment of my arguments in this section are largely a rehash of standard critiques of the "explanatory gap" (Levine 1983) and the hard problem of consciousness (Chalmers 1996). The standard refrain is that the inability of science to account for subjectivity corresponds to an epistemological gap not an ontological one. The fact that we cannot explain something, and perhaps never will be able to, does not require us to assume a different ontological foundation for reality. I concede that this kind of mysterian (McGinn 1989) account of the explanatory gap, although profoundly unsatisfying, is difficult to dispute. However, she largely ignores the more novel aspects of my arguments. Namely, she disregards my claim that not only is the current physicalist account unable to explain consciousness, it outright rejects two additional subjectively self-evident aspects of reality. It rejects the flow of time and the privileged present. While she acknowledges in a footnote that she finds this aspect of the paper "inspiring," it does not impact her overall dismissal of the need for a meta-perspective. As she puts it, "what the proposed meta-perspective might be and how it is helpful despite acknowledging our common sense intuition eludes my understanding not at an epistemological level but at an ontological level" (Gottschling this collection, p. 23).

Gottschling's reaction to the third section of my paper was not unexpected. As I noted in the close of my paper, "my arguments on this point will likely remain wholly unpersuasive to those who cannot conceive of subjective experience as offering an epistemological authority that rivals science." I recognize that it will be an uphill

battle to persuade philosophers and scientists steeped in the supremacy of the third-person perspective to consider that conclusions drawn from our own experience could possibly carry ramifications comparable to conventional objective science. But at the end of the day all of the science that we believe we know is necessarily delivered to us through our subjective experience. While what we know about objective reality is necessarily dependent on experience, the same is not the case for experience. Objective reality could conceivably be an illusion. This could all be a dream or we could be the proverbial brain in a vat. But the *experience* of objective reality is unquestionable, as even an illusory experience is still an experience. Given that the existence of objective reality is ultimately on less certain ground than the existence of experience, it is far from obvious why the third-person frame-of-reference holds its current unchallenged dominion.

5 Conclusion

I suspect that my big-picture approach to replying to Gottschling's very detailed analysis may be unsatisfying to some (Gottschling included) who might have expected point-by-point replies to each of her concerns. However, I hope that my stepping-back tactic enabled me to address the major concerns that were raised. At the outset I noted the close parallels between the factors that contribute to conceptual and perceptual processes. In addition to the value of perspective shifting, it might also be noted that stepping-back is another strategy that is useful in both conceptual and perceptual domains. For example, it is easier to decipher a highly pixelated photo from a distance than up close. Similarly, when people confront conceptual insight problems from a more distant perspective (e.g., imagining themselves a year from now) they are often better able reach a solution (Förster et al. 2004). Conceptual stepping back can enable one to distinguish the metaphorical "forest from the trees." It remains unclear whether there could be a genuine meta-perspective that enables us to accommodate the assumptions of both the first- and third- person perspectives. But if such a perspective does ex-

ist, it seems likely that finding it will require stepping back...way back.

Acknowledgements

I thank Ashley Brumett and Tam Hunt for comments on an earlier version of this paper. The writing of this reply was supported by a grant from the Institute of Educational Science grant R305A110277. The content of this reply does not necessarily reflect the position or policy of the US government, and no official endorsement should be inferred.

References

- Baer, R. A., Smith, G. T., Hopkins, J., Krietemeyer, J. & Toney, L. (2006). Using self-report assessment methods to explore facets of mindfulness. *Assessment*, 13 (1), 27-45. [10.1177/1073191105283504](https://doi.org/10.1177/1073191105283504)
- Block, N. (1995). On a confusion about a function of consciousness. *Behavioural and Brain Sciences*, 18, 227-287.
- Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. New York, NY: Oxford University Press.
- Chambers, D. & Reisberg, D. (1992). What an image depicts depends on what an image means. *Cognitive Psychology*, 24, 145-174.
- Dehaene, S., Changeux, J. P., Naccache, L., Sackur, J. & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: A testable taxonomy. *Trends in Cognitive Sciences*, 10, 204-211.
- Einstein, A. (2001). *Relativity: The special and the general theory (reprint of 1920 translation by Robert W. Lawson ed.)*. London, UK: Routledge.
- Förster, J., Friedman, R. S. & Liberman, N. (2004). Temporal construal effects on abstract and concrete thinking: Consequences for insight and creative cognition. *Journal of personality and social psychology*, 87 (2), 177-189.
- Gottschling, V. (2015). Bridging the gap-A commentary on Jonathan W. Schooler. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, 64, 354-361.
- Lutz, A., Slagter, H. A., Dunne, J. D. & Davidson, R. J. (2008). Attention regulation and monitoring in meditation. *Trends in Cognitive Sciences*, 12 (4), 163-169. [10.1016/j.tics.2008.01.005](https://doi.org/10.1016/j.tics.2008.01.005)
- McGinn, C. (1989). Can we solve the mind-body problem? *Mind*, 98 (391), 349-366.
- Rosenthal, D. (1986). Two concepts of consciousness. *Philosophical Studies*, 49 (3), 329-359. [10.1007/BF00355521](https://doi.org/10.1007/BF00355521)
- Schooler, J. W. (2002). Re-representing consciousness: Dissociations between experience and meta-consciousness. *Trends in Cognitive Sciences*, 6 (8), 339-344. [10.1016/S1364-6613\(02\)01949-6](https://doi.org/10.1016/S1364-6613(02)01949-6)
- (2015). Bridging the objective/subjective divide. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Schooler, J. W., Smallwood, J., Christoff, K., Handy, T. C., Reichle, E. D. & Sayette, M. A. (2011). Meta-awareness, perceptual decoupling and the wandering mind. *Trends in Cognitive Sciences*, 15, 319-326. [doi/org/10.1016/j.tics.2011.05.006](https://doi.org/doi/org/10.1016/j.tics.2011.05.006)
- Schooler, J. W., Mrazek, M. D., Baird, B. & Winkielman, P. (2015). Minding the mind: The value of distinguishing among unconscious, conscious, and metaconscious processes. In M. Mikulincer, P. R. Shaver, E. Borgida & J. A. Bargh (Eds.) *APA handbook of personality and social psychology, Vol. 1. Attitudes and social cognition* (pp. 179-202). Washington, DC: American Psychological Association.
- Schooler, J. W. & Melcher, J. (1995). The ineffability of insight. In S. M. Smith, T. B. Ward & R. A. Finke (Eds.) *The creative cognition approach* (pp. 97-134). Cambridge, MA: MIT Press.
- Seli, P., Jonker, T. R., Cheyne, J. A., Cortes, K. & Smilek, D. (in press). Can research participants comment authoritatively on the validity of their self-reports of mind wandering and task engagement? *Journal of Experimental Psychology: Human Perception and Performance*
- Thompson, E. (2014). *Waking, dreaming, being: Self and consciousness in neuroscience, meditation, and philosophy*. New York, NY: Columbia University Press.
- Tversky, A. & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211 (4481), 453-458. [10.1126/science.7455683](https://doi.org/10.1126/science.7455683)
- Westfall, R. S. (1980). *Never at rest: A biography of Isaac Newton*. Cambridge, UK: Cambridge University Press.
- Wiseman, R., Watt, C., Gilhooly, K. & Georgiou, G. (2011). Creativity and ease of ambiguous figural reversal. *British Journal of Psychology*, 102 (3), 615-622. [10.1111/j.2044-8295.2011.02031.x](https://doi.org/10.1111/j.2044-8295.2011.02031.x)

The Cybernetic Bayesian Brain

From Interoceptive Inference to Sensorimotor Contingencies

Anil K. Seth

Is there a single principle by which neural operations can account for perception, cognition, action, and even consciousness? A strong candidate is now taking shape in the form of “predictive processing”. On this theory, brains engage in predictive inference on the causes of sensory inputs by continuous minimization of prediction errors or informational “free energy”. Predictive processing can account, supposedly, not only for perception, but also for action and for the essential contribution of the body and environment in structuring sensorimotor interactions. In this paper I draw together some recent developments within predictive processing that involve predictive modelling of internal physiological states (*interoceptive inference*), and integration with “enactive” and “embodied” approaches to cognitive science (*predictive perception of sensorimotor contingencies*). The upshot is a development of predictive processing that originates, not in Helmholtzian perception-as-inference, but rather in 20th-century cybernetic principles that emphasized homeostasis and predictive control. This way of thinking leads to (i) a new view of emotion as active interoceptive inference; (ii) a common predictive framework linking experiences of body ownership, emotion, and exteroceptive perception; (iii) distinct interpretations of active inference as involving disruptive and disambiguatory—not just confirmatory—actions to test perceptual hypotheses; (iv) a neurocognitive operationalization of the “mastery of sensorimotor contingencies” (where sensorimotor contingencies reflect the rules governing sensory changes produced by various actions); and (v) an account of the sense of subjective reality of perceptual contents (“perceptual presence”) in terms of the extent to which predictive models encode potential sensorimotor relations (this being “counterfactual richness”). This is rich and varied territory, and surveying its landmarks emphasizes the need for experimental tests of its key contributions.

Keywords

Active inference | Counterfactually-equipped predictive model | Evolutionary robotics | Free energy principle | Interoception | Perceptual presence | Predictive processing | Sensorimotor contingencies | Somatic marker hypothesis | Synaesthesia

1 Introduction

An increasingly popular theory in cognitive science claims that brains are essentially prediction machines (Hohwy 2013). The theory is variously known as the Bayesian brain (Knill & Pouget 2004; Pouget et al. 2013), predictive processing (Clark 2013; Clark this collection), and the predictive mind (Hohwy 2013; Hohwy this collection), among others; here we use the term PP (predictive processing). (See Table 1 for a glossary of technical terms.) At its most fundamental, PP says that perception is the res-

ult of the brain inferring the most likely causes of its sensory inputs by minimizing the difference between actual sensory signals and the signals expected on the basis of continuously updated predictive models. Arguably, PP provides the most complete framework to date for explaining perception, cognition, and action in terms of fundamental theoretical principles and neurocognitive architectures. In this paper I describe a version of PP that is distinguished by (i) an emphasis on predictive modelling of in-

Author

Anil K. Seth

a.k.seth@sussex.ac.uk

University of Sussex

Brighton, United Kingdom

Commentator

Wanja Wiese

wawiese@uni-mainz.de

Johannes Gutenberg-Universität

Mainz, Germany

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität

Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University

Melbourne, Australia

Table 1: A glossary of technical terms.

Allostasis	The process of achieving homeostasis.
Active inference	Classically conceived as the minimization of prediction error by performing actions that confirm sensory predictions. However, as argued in this paper, it may also involve the performance of actions to disconfirm current predictions or to disambiguate among competing perceptual hypotheses.
Counterfactually-equipped predictive model	A predictive or generative model that encodes not only the likely causes of current sensory inputs but also (and explicitly) the likely causes of fictive sensory inputs conditioned on possible but unexecuted actions.
Counterfactual richness	A predictive model is counterfactually rich if it encodes a rich repertoire of potential sensorimotor relations, i.e., relations between potential actions and their expected sensory consequences.
Exteroception/exteroceptive	The classic senses conveying signals originating in the external environment.
Free energy	An information-theoretic quantity that bounds or limits the surprise associated with encountering an input, given a generative/predictive model mapping causes to sensory inputs. Under fairly general assumptions, free energy is the long-run sum of prediction error.
Free energy principle (FEP)	The FEP says that organisms obey a fundamental imperative towards the avoidance of (information-theoretically) surprising events, according to which they must minimize the long-run average surprise of sensory states, since surprising sensory states are (in the long run) likely to reflect conditions incompatible with continued existence.
Homeostasis	Any regulative processes that enable a system to keep certain variables within specific bounds.
Interoception/interoceptive	The sense of the internal physiological condition of the body.
Interoceptive inference	The predictive modelling of internal physiological states.
Interoceptive sensitivity	A characterological trait that reflects individual sensitivity to interoceptive signals, usually operationalized via heartbeat detection tasks.
Perceptual presence	The sense of the subjective reality of the contents of perception.
PPSMC	Predictive Perception of SensoriMotor Contingencies. A new theory that integrates predictive processing with sensorimotor theory. It says that mastery of a sensorimotor contingency is equivalent to the induction and deployment of a counterfactually-equipped predictive model linking potential actions to their expected sensory consequences.
Predictive processing (PP)/predictive coding	A scheme, dating back at least to Hermann von Helmholtz, which conceives of perception as probabilistic inference on the causes of sensory signals. Predictive coding is one specific implementation of predictive processing that rests on algorithms developed in the setting of data compression.
Sensorimotor contingency (SMC)	SMCs describe ways in which sensory signals change given actions in specific contexts; they are “rules” describing sensorimotor dependencies.
Sensorimotor theory	A cognitive theory which says that visual experiences arises from an implicit knowledge or mastery of SMCs. On this theory, perception is an activity.

ternal physiological states and (ii) engagement with alternative frameworks under the banner of “enactive” and “embodied” cognitive science (Varela et al. 1993).

I first identify an unusual starting point for PP, not in Helmholtzian perception-as-inference, but in the mid 20th-century cybernetic theories associated with W. Ross Ashby (1952, 1956; Conant & Ashby 1970). Linking these origins to their modern expression in Karl Friston’s “free energy principle” (2010), perception emerges as a *consequence* of a more fundamental imperative towards homeostasis and control, and not as a process designed to furnish a detailed inner “world model” suitable for cognition and action planning. The ensuing view of PP, while still fluently accounting for (exteroceptive) perception, turns out to be more naturally applicable to the predictive perception of internal bodily states, instantiating a process of *interoceptive inference* (Seth 2013; Seth et al. 2011). This concept provides a natural way of thinking of the neural substrates of emotional and mood experiences, and also describes a common mechanism by which interoceptive and exteroceptive signals can be integrated to provide a unified experience of body ownership and conscious selfhood (Blanke & Metzinger 2009; Limanowski & Blankenburg 2013).

The focus on embodiment leads to distinct interpretations of *active inference*, which in general refers to the selective sampling of sensory signals so as to improve perceptual predictions. The simplest interpretation of active inference is the changing of sensory data (via selective sampling) to conform to current predictions (Friston et al. 2010). However, by analogy with hypothesis testing in science, active inference can also involve seeking evidence that goes *against* current predictions, or that *disambiguates* multiple competing hypotheses. A nice example of the latter comes from self-modelling in evolutionary robotics, where multiple competing self-models are used to specify actions that are most likely to provide disambiguatory sensory evidence (Bongard et al. 2006). I will spend more time on this example later. Crucially, these different senses of active inference rest on the capacity of predictive models to encode

counterfactual relations linking potential (but not necessarily executed) actions to their expected sensory consequences (Friston et al. 2012; Seth 2014b). It also implies the involvement of model comparison and selection—not just the optimization of parameters assuming a single model. These points represent significant developments in the basic infrastructure of PP.

The notion of counterfactual predictions connects PP with what at first glance seems to be its natural opponent: “enactive” theories of perception and cognition that explicitly reject internal models or representations (Clark this collection; Hutto & Myin 2013; Thompson & Varela 2001). Central to the enactive approach are notions of “sensorimotor contingencies” and their “mastery” (O’Regan & Noë 2001), where a sensorimotor contingency refers to a rule governing how sensory signals change in response to action. On this approach, the perceptual experience of (for example) redness is given by an implicit knowledge (mastery) of the way red things behave given certain patterns of sensorimotor activity. This mastery of sensorimotor contingencies is also said to underpin *perceptual presence*: the sense of subjective reality of the contents of perception (Noë 2006). From the perspective of PP, mastery of a sensorimotor contingency corresponds to the learning of a counterfactually-equipped predictive model connecting potential actions to expected sensory consequences. The resulting theory of PPSMC (Predictive Perception of SensoriMotor Contingencies), Seth 2014b) provides a much needed reconciliation of enactive and predictive theories of perception and action. It also provides a solution to the challenge of perceptual presence within the setting of PP: perceptual presence obtains when the underlying predictive models are *counterfactually rich*, in the sense of encoding a rich repertoire of potential (but not necessarily executed) sensorimotor relations. This approach also helps explain instances where perceptual presence seems to be lacking, such as in synaesthesia.

This is both a conceptual and theoretical paper. Space limitations preclude any significant treatment of the relevant experimental lit-

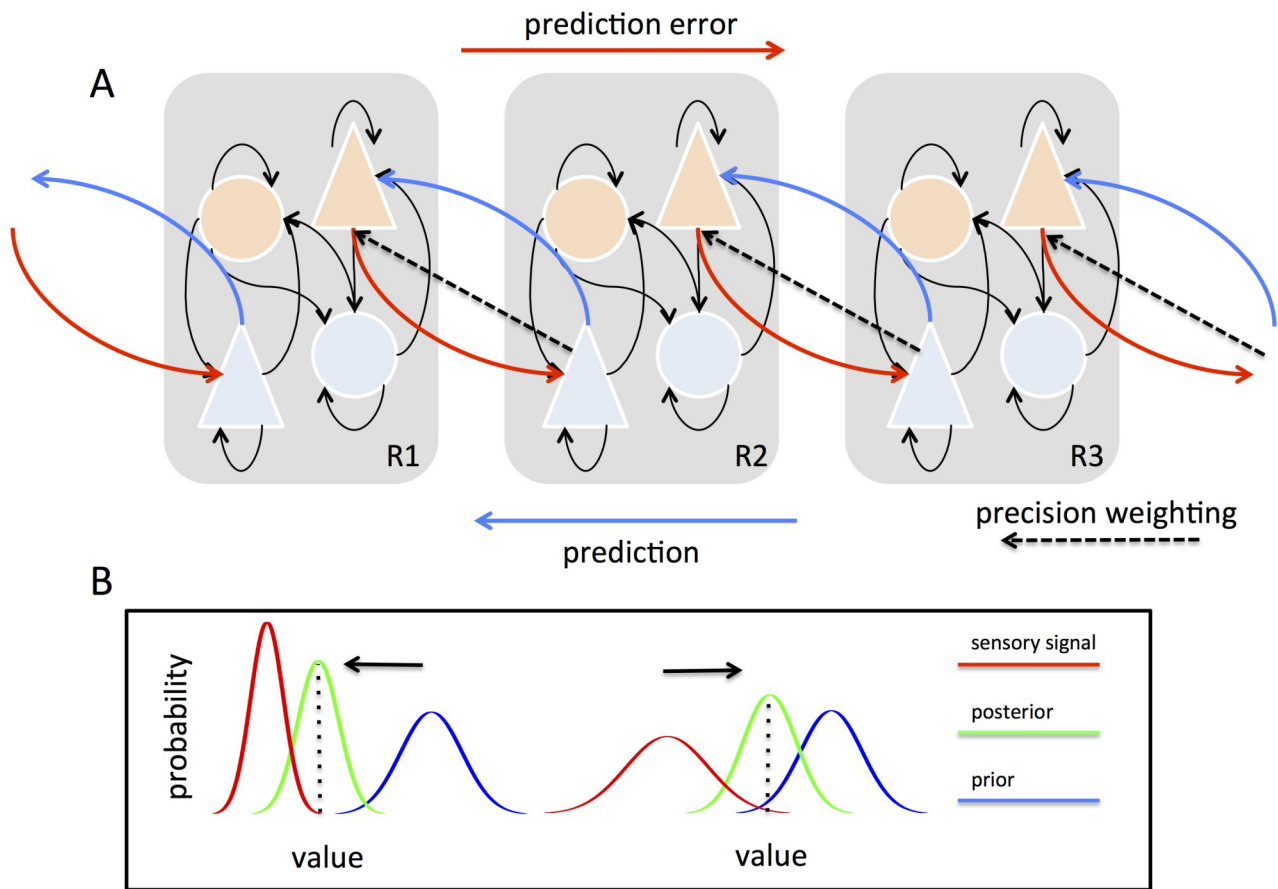


Figure 1: **A.** Schemas of hierarchical predictive coding across three cortical regions; the lowest on the left (R1) and the highest on the right (R3). Bottom-up projections (red) originate from “error units” (orange) in superficial cortical layers and terminate on “state units” (light blue) in the deep (infragranular) layers of their targets; while top-down projections (dark blue) convey predictions originating in deep layers and project to the superficial layers of their targets. Prediction errors are associated with precisions, which determine the relative influence of bottom-up and top-down signal flow via precision weighting (dashed lines). **B.** The influence of precisions on Bayesian inference and predictive coding. The curves show probability distributions over the value of a sensory signal (x -axis). On the left, high precision-weighting of sensory signals (red) enhances their influence on the posterior (green) and expectation (dotted line) as compared to the prior (blue). On the right, low sensory precision weighting has the opposite effect. Figure adapted from Seth (2013).

erature. However, even an exhaustive treatment would reveal that this literature so far provides only circumstantial support for the basics of PP, let alone for the extensions described here. Yet an advantage of PP theories is that they are grounded in concrete computational processes and neurocognitive architectures, giving us confidence that informative experimental tests can be devised. Implementing such an experimental agenda stands as a critical challenge for the future.

2 The predictive brain and its cybernetic origins

2.1 Predictive processing: The basics

PP starts with the assumption that in order to support adaptive responses, the brain must discover information about the external “hidden” causes of sensory signals. It lacks any direct access to these causes, and can only use information found in the flux of sensory signals them-

selves. According to PP, brains meet this challenge by attempting to predict sensory inputs on the basis of their own emerging models of the causes of these inputs, with prediction errors being used to update these models so as to minimize discrepancies. The idea is that a brain operating this way will come to encode (in the form of predictive or generative models) a rich body of information about the sources of signals by which it is regularly perturbed (Clark 2013).

Applied to cortical hierarchies, PP overturns classical notions of perception that describe a largely “bottom-up” process of evidence accumulation or feature detection. Instead, PP proposes that perceptual content is determined by top-down predictive signals emerging from multi-layered and hierarchically-organized generative models of the causes of sensory signals (Lee & Mumford 2003). These models are continually refined by mismatches (prediction errors) between predicted signals and actual signals across hierarchical levels, which iteratively update predictive models via approximations to Bayesian inference (see Figure 1). This means that the brain can induce accurate generative models of environmental hidden causes by operating only on signals to which it has direct access: *predictions* and *prediction errors*. It also means that even low-level perceptual content is determined via cascades of predictions flowing from very general abstract expectations, which constrain successively more fine-grained predictions.

Two further aspects of PP need to be emphasized from the outset. First, sensory prediction errors can be minimized either “passively”, by changing predictive models to fit incoming data (perceptual inference), or “actively”, by performing actions to confirm or test sensory predictions (active inference). In most cases these processes are assumed to unfold continuously and simultaneously, underlining a deep continuity between perception and action (Friston et al. 2010; Verschure et al. 2003). This process of active inference will play a key role in much of what follows. Second, predictions and prediction errors in a Bayesian framework have associated *precisions* (inverse variances, Figure 1). The precision of a prediction error is an in-

dicator of its reliability, and hence can be used to determine its influence in updating top-down predictive models. Precisions, like mean values, are not given but must be inferred on the basis of top-down models and incoming data; so PP requires that agents have *expectations about precisions* that are themselves updated as new data arrive (and new precisions can be estimated). Precision expectations can therefore balance the influence of different prediction-error sources on the updating of predictive models. And if prediction errors have low (expected) precision, predictive models may overwhelm error signals (hallucination) or elicit actions that confirm sensory predictions (active inference).

A picture emerges in which cortical networks engage in recurrent interactions whereby bottom-up prediction errors are continuously reconciled with top-down predictions at multiple hierarchical levels—a process modulated at all times by precision weighting. The result is a brain that not only encodes information about the sources of signals that impinge upon its sensory surfaces, but that also encodes information about how its own actions interact with these sources in specifying sensory signals. *Perception* involves updating the parameters of the model to fit the data; *action* involves changing sensory data to fit (or test) the model; and *attention* corresponds to optimizing model updating by giving preference to sensory data that are expected to carry more information, which is called precision weighting (Hohwy 2013). This view of the brain is shamelessly model-based and representational (though with a finessed notion of representation), yet it also deeply embeds the close coupling of perception and action and, as we will see, the importance of the body in the mediation of this interaction.

2.2 Predictive processing and the free energy principle

PP can be considered a special case of the *free energy principle*, according to which perceptual inference and action emerge as a consequence of a more fundamental imperative towards the avoidance of “surprising” events (Friston 2005, 2009, 2010). On the free energy principle, or-

ganisms – by dint of their continued survival—must minimize the long-run average surprise of sensory states, since surprising sensory states are likely to reflect conditions incompatible with continued existence (think of a fish out of water). “Surprise” is not used here in the psychological sense, but in an information-theoretic sense—as the negative log probability of an event’s occurrence (roughly, the unlikeliness of the occurrence of an event).

The connection with PP arises because agents cannot directly evaluate the (information-theoretic) surprise associated with an event, since this would require—impossibly—the agent to average over all possible occurrences of the event in all possible situations. Instead, the agent can only maintain a lower limit on surprise by minimizing the difference between actual sensory signals and those signals predicted according to a generative or predictive model. This difference is *free energy*, which, under fairly general assumptions, is the long-run sum of prediction error.

An attractive feature of the free energy principle is that it brings to the table a rich mathematical framework that shows how PP can work in practice. Formally, PP depends on established principles of Bayesian inference and model specification, whereby the most likely causes of observed data (*posterior*) are estimated based on optimally combining *prior expectations* of these causes with observed data, by using a (generative, predictive) model of the data that would be observed given a particular set of causes (*likelihood*). (See Figure 1 for an example of priors and posteriors.) In practice, because optimal Bayesian inference is usually intractable, a variety of approximate methods can be applied (Hinton & Dayan 1996; Neal & Hinton 1998). Friston’s framework appeals to previously worked-out “variational” methods, which take advantage of certain approximations (e.g., Gaussianity, independence of temporal scales)—thus allowing a potentially neat mapping onto neurobiological quantities (Friston et al. 2006).¹

1 Some challenging questions surface here as to whether prediction errors are used to update priors, which corresponds to standard Bayesian inference, or whether they are used to update the underlying generative/predictive model, which corresponds to learning.

The free energy principle also emphasizes *action* as a means of prediction error minimization, this being *active inference*. In general, active inference involves the selective sampling of sensory signals so as to minimize uncertainty in perceptual hypotheses (minimizing the entropy of the posterior). In one sense this means that actions are selected to provide evidence compatible with current perceptual predictions. This is the most standard interpretation of the concept, since it corresponds most directly to minimization of prediction error (Friston 2009). However, as we will see, actions can also be selected on the basis of an attempt to find evidence going against current hypotheses, and/or to efficiently disambiguate between competing hypotheses. These finessed senses of active inference represent developments of the free energy framework. Importantly, action itself can be thought of as being brought about by the minimization of *proprioceptive* prediction errors via the engagement of classical reflex arcs (Adams et al. 2013; Friston et al. 2010). This requires transiently low precision-weighting of these errors (or else predictions would simply be updated instead), which is compatible with evidence showing sensory attenuation during self-generated movements (Brown et al. 2013).

A more controversial aspect of the free energy principle is its claimed generality (Hohwy this collection). At least as described by Friston, it claims to account for adaptation at almost any granularity of time and space, from macroscopic trends in evolution, through development and maturation, to signalling in neuronal hierarchies (Friston 2010). However, in some of these interpretations reliance on predictive modelling is only implicit; for example the body of a fish can be considered to be an implicit model of the fluid dynamics and other affordances of its watery environment (see section 2.3). I am not concerned here with these broader interpretations, but will focus on those cases in which biological (neural) mechanisms plausibly implement explicit predictive inference via approximations to Bayesian computations—namely, the Bayesian brain (Knill & Pouget 2004; Pouget et al. 2013). Here, the free energy principle has potentially the greatest explanat-

ory power, especially given the convergence of empirical evidence (see [Clark 2013](#) and [Hohwy 2013](#) for reviews) and computational modelling showing how cortical microcircuits might implement approximate Bayesian inference ([Bastos et al. 2012](#)).

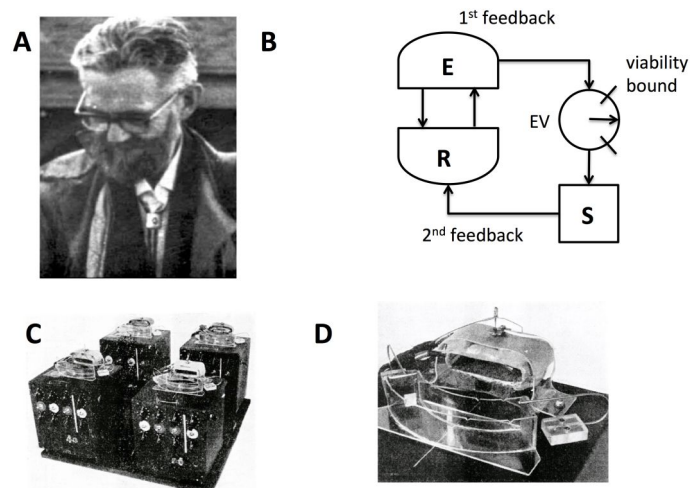


Figure 2: **A.** W. Ross Ashby, British psychiatrist and pioneer of cybernetics (1903–1972). **B.** A schematic of ultrastability, based on Ashby’s notebooks. The system R homeostatically maintains its essential variables (EVs) within viability limits via first-order feedback with the environment E . When first-order feedback fails, so that EVs run out-of-bounds, second order “ultrastable” feedback is triggered so that S (an internal controller, potentially model-based) changes the parameters of R governing the first-order feedback. S continually changes R until homeostatic relations are regained, leaving the EVs again within bounds. **C.** Ashby’s “homeostat”, consisting of four interconnected ultrastable systems, forming a so-called “multistable” system. **D.** One ultrastable unit from the homeostat. Each unit had a trough of water with an electric field gradient and a metal needle. Instability was represented by the non-central needle positions, which on occurring would alter the resistances connecting the units via discharge through capacitors. For more details see [Ashby \(1952\)](#) and [Pickering \(2010\)](#).

2.3 Predictive processing, free energy, and cybernetics

Typically, the origins of PP are traced to the work of the 19th Century physiologist Hermann von Helmholtz, who first formalized the idea of perception as inference. However, the Helmholt-

zian view is rather passive, inasmuch as there is little discussion of active inference or behaviour. The close coupling of perception and action emphasized in the free energy principle points instead to a deep connection between PP and mid-twentieth-century cybernetics. This is most obvious in the works of [W. Ross Ashby \(Ashby 1952; 1956; Conant & Ashby 1970\)](#) but is also evident more generally ([Dupuy 2009; Pickering 2010](#)). Importantly, cybernetics adopted as its central focus the *prediction and control of behaviour* in so-called teleological or purposeful machines.² More precisely, cybernetic theorists were (are) interested in systems that appear to have goals (i.e., teleological) and that participate in circular causal chains (i.e., involving feedback) coupling goal-directed sensation and action.

Two key insights from the first wave of cybernetics usefully anticipate the core developments of PP within cognitive science. These are both associated with Ashby, a key figure in the movement and often considered its leader, at least outside the USA ([Figure 2](#)).

The first insight consists in an emphasis on the homeostasis of internal *essential variables*, which, in physiological settings, correspond to quantities like blood pressure, heart rate, blood sugar levels, and the like. In Ashby’s framework, when essential variables move beyond specific viability limits, adaptive processes are triggered that re-parameterize the system until it reaches a new equilibrium in which homeostasis is restored ([Ashby 1952](#)). Such systems are, in Ashby’s terminology, *ultrastable*, since they embody (at least) two levels of feedback: a first-order feedback that homeostatically regulates essential variables (like a thermostat) and a second-order feedback that allostatically³ re-organises a system’s input–output relations when first-order feedback fails, until a new homeostatic regime is attained. In the most basic case, as implemented in Ashby’s famous “homeostat” ([Figure 2](#)), this second-order feedback simply involves random changes to system

² This underlines the close links between cybernetics and behaviourism. Perhaps this explains why cybernetics was so reluctant to bring phenomenology into its remit, an exclusion which, looking back, seems like a missed opportunity.

³ Allostasis: the process of achieving homeostasis.

parameters until a new stable regime is reached. The importance of this insight for PP is that it locates the function of biological and cognitive processes in generalizing homeostasis to ensure that internal essential variables remain within expected ranges.

Another way to summarize the fundamental cybernetic principle is to say that adaptive systems ensure their continued existence by successfully responding to environmental perturbations so as to maintain their internal organization. This leads to the second insight, evident in Ashby's *law of requisite variety*. This states that a successful control system must be capable of entering at least as many states as the system being controlled: "only variety can force down variety" (Ashby 1956). This induces a functional boundary between controller and environment and implies a minimum level of complexity for a successful controller, which is determined by the causal complexity of the environmental states that constitute potential perturbations to a system's essential variables. This view was refined some years later, in a 1970 paper written with Roger Conant entitled "Every good regulator of a system must be a model of that system" (Conant & Ashby 1970). This paper builds on the law of requisite variety by arguing (and attempting to formally show) that the nature of a controller capable of suppressing perturbations imposed by an external system (e.g., the world) must instantiate a model of that system. This provides a clear connection with the free energy principle, which proposes that adaptive systems minimize a limit on free energy (long-run average surprise) by inducing and refining a generative model of the causes of sensory signals. It also moves beyond Ashby's homeostat by implying that model-based controllers can engage in more successful multi-level feedback than is possible by random variation of higher-order parameters.

Putting these insights together provides a distinctive way of seeing the relevance of PP to cognition and biological adaptation. It can be summarized as follows. The purpose of cognition (including perception and action) is to maintain the homeostasis of essential variables and of internal organization (ultrastability).

This implies the existence of a control mechanism with sufficient complexity to respond to (i.e., suppress) the variety of perturbations it encounters (law of requisite variety). Further, this structure must instantiate a model of the system to be controlled (good regulator theorem), where the system includes both the body and the environment (and their interactions). As Ashby himself tells us "[t]he whole function of the brain can be summed up in: error correction" (quoted in Clark 2013, p. 1). Put this way, perception emerges as a *consequence* of a more fundamental imperative towards organizational homeostasis, and not as a stage in some process of internal world-model construction. This view, while highlighting different origins, closely parallels the assumptions of the free energy principle in proposing a primary imperative towards the continued survival of the organism (Friston 2010).

It may be surprising to consider the legacy of cybernetics in this light. This is because many previous discussions of this legacy focus on examples which show that complex, apparently goal-directed behaviour can emerge from simple mechanisms interacting with structured bodies and environments (Beer 2003; Braitenberg 1984). On this more standard development, cybernetics challenges rather than asserts the need for internal models and representations: it is often taken to justify slogans of the sort "the world is its own best model" (Brooks 1991). In fact, cybernetics is agnostic with respect to the need for deployment of explicit internally-specified predictive models. If environmental circumstances are reasonably stable, and mappings between perturbations and (homeostatic) responses reasonably straightforward, then the good regulator theorem can be satisfied by controllers that only implicitly model their environments. This is the case, for instance, in the Watt governor: a device that is able exquisitely to control the output of (for instance) a steam engine, in virtue of its mechanism, and not through the deployment of explicit predictive models or representations (see Figure 3 and Van Gelder 1995; note that the governor can

be described as an implicit model since it has variables – e.g., eccentricity of the metal balls from the central column – which map onto environmental variables that affect the homeostatic target – engine output). However, where there exist many-to-many mappings between sensory states and their probable causes, as may be the case more often than not, it will pay to engage explicit inferential processes in order to extract the most probable causes of sensory states, insofar as these causes threaten the homeostasis of essential variables.

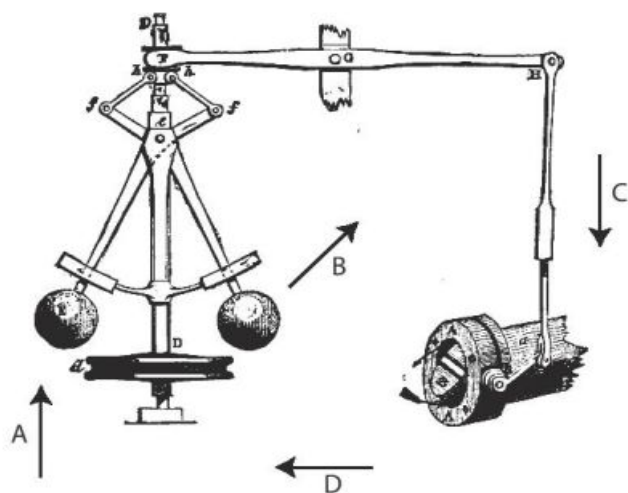


Figure 3: The Watt governor. This system, a central contributor to the industrial revolution, enabled precise control over the output of (for example) steam engines. As the speed of the engine increases, power is supplied to the governor (A) by a belt or chain, causing it to rotate more rapidly so that the metal balls have more kinetic energy. This causes the balls to rise (B), which closes the throttle valve (C), thereby reducing the steam flow, which in turn reduces engine speed (D). The opposite happens when the engine speed decreases, so that the governor maintains engine speed at a precise equilibrium.

In summary, rather than seeing PP as originating solely in the Helmholtzian notion of “perception as inference”, it is fruitful to see it also as part of a process of model-based *predictive control* entailed by a fundamental imperative towards internal homeostasis. This shift in perspective reveals a distinctive agenda for PP in cognitive science, to which I shall now turn.

3 Interoceptive inference, emotion, and predictive selfhood

3.1 Interoceptive inference and emotion

Considering the cybernetic roots of PP, together with the free energy principle, leads to a potentially counterintuitive idea. This is that PP may apply more naturally to *interoception* (the sense of the internal physiological condition of the body) than to *exteroception* (the classic senses, which carry signals that originate in the external environment). This is because for an organism it is more important to avoid encountering unexpected interoceptive states than to avoid encountering unexpected exteroceptive states. A level of blood oxygenation or blood sugar that is unexpected is likely to be bad news for an organism, whereas unexpected exteroceptive sensations (like novel visual inputs) are less likely to be harmful and may in some cases be desirable, as organisms navigate a delicate balance between exploration and exploitation (Seth 2014a), testing current perceptual hypotheses through active inference (see section 5, below), all ultimately in the service of maintaining organismic homeostasis.

Perhaps because of its roots in Helmholtz, PP has largely been developed in the setting of visual neuroscience (Rao & Ballard 1999), with a related but somewhat independent line in motor control (Wolpert & Ghahramani 2000). Recently, an explicit application of PP to interoception has been developed (Seth 2013; Seth & Critchley 2013; Seth et al. 2011; see also Gu et al. 2013). On this theory of *interoceptive inference* (or equivalently *interoceptive predictive coding*), emotional states (i.e., subjective feeling states) arise from top-down predictive inference of the causes of interoceptive sensory signals (see Figure 4). In direct analogy to exteroceptive PP, emotional content is constitutively specified by the content of top-down interoceptive predictions *at a given time*, marking a distinction with the well-studied impact of expectations on *subsequent* emotional states (see e.g., Ploghaus et al. 1999; Ueda et al. 2003). Furthermore, interoceptive prediction errors can

be minimized by (i) updating predictive models (perception, corresponding to new emotional contents); (ii) changing interoceptive signals through engaging autonomic reflexes (autonomic control or active inference); or (iii) performing behaviour so as to alter external conditions that impact on internal homeostasis (allostasis; Gu & Fitzgerald 2014; Seth et al. 2011).

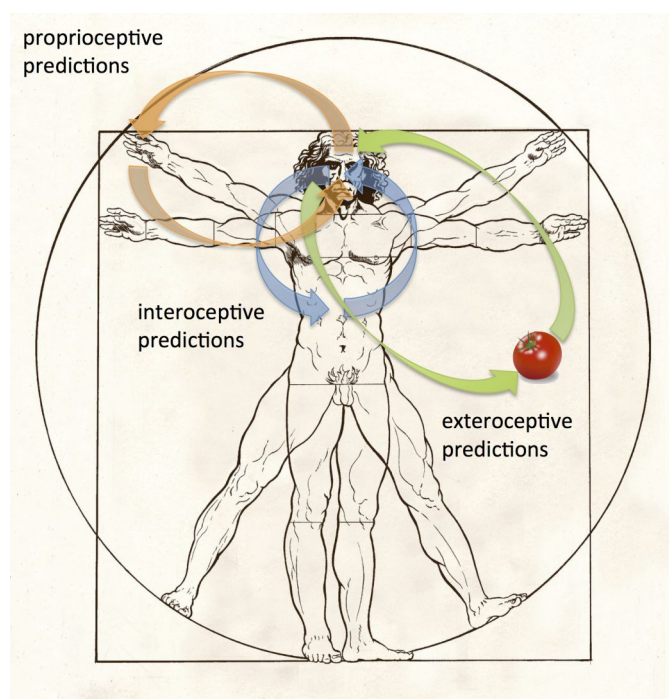


Figure 4: Inference and perception. Green arrows represent exteroceptive predictions and predictions errors underpinning perceptual content, such as the visual experience of a tomato. Orange arrows represent proprioceptive predictions (and prediction errors) underlying action and the experience of body ownership. Blue arrows represent interoceptive predictions (and prediction errors) underlying emotion, mood, and autonomic regulation. Hierarchically higher levels will deploy multimodal and even amodal predictive models spanning these domains, which are capable of generating multimodal predictions of afferent signals.

Consider an example in which blood sugar levels (an essential variable) fall towards or beyond viability thresholds, reaching unexpected and undesirable values (Gu & Fitzgerald 2014; Seth et al. 2011). Under interoceptive inference, the following responses ensue. First, interoceptive prediction error signals update top-down expectations, leading to sub-

jective experiences of hunger or thirst (for sugary things). Because these feeling states are themselves surprising (and non-viable) in the long run, they signal prediction errors at hierarchically-higher levels, where predictive models integrate multimodal interoceptive and exteroceptive signals. These models instantiate predictions of temporal sequences of matched exteroceptive and interoceptive inputs, which flow down through the hierarchy. The resulting cascade of prediction errors can then be resolved either through autonomic control, in order to metabolize bodily fat stores (active inference), or through allostatic actions involving the external environment (i.e., finding and eating sugary things).

The sequencing and balance of these events is governed by relative precisions and their expectations. Initially, interoceptive prediction errors have high precision (weighting) given a higher-level expectation of stable homeostasis. Whether the resulting high-level prediction error engages autonomic control or allostatic behaviour (or both) depends on the precision weighting of the corresponding prediction errors. If food is readily available, consummatory actions lead to food intake (as described earlier, these actions are generated by the resolution of proprioceptive prediction errors). If not, autonomic reflexes initiate the metabolization of bodily fat stores, perhaps alongside appetitive behaviours that are predicted to lead to the availability of food, conditioned on performing these behaviours.⁴

3.2 Implications of interoceptive inference

Several interesting implications arise when considering emotion as resulting from interoceptive inference (Seth 2013). First, the theory generalizes previous “two factor” theories of emotion that see emotional content as resulting from an interaction between the perception of physiolo-

⁴ It is interesting to consider possible dysfunctions in this process. For example, if high-level predictions about the persistence of low blood sugar become abnormally strong (i.e., low blood sugar becomes chronically expected), allostatic food-seeking behaviours may not occur. This process, akin to the transition from hallucination to delusion in perceptual inference (Fletcher & Frith 2009), may help understand eating disorders in terms of dysfunctional signalling of satiety.

gical changes (James 1894) and “higher-level” cognitive appraisal of the context within which these changes occur (Schachter & Singer 1962). Instead of distinguishing “physiological” and “cognitive” levels of description, interoceptive inference sees emotional content as resulting from the multi-layered prediction of interoceptive input spanning many levels of abstraction. Thus, interoceptive inference integrates cognition and emotion within the powerful setting of PP.

The theory also connects with influential frameworks that link interoception with decision making, notably the “somatic marker hypothesis” proposed by Antonio Damasio (1994). According to the somatic marker hypothesis, intuitive decisions are shaped by interoceptive responses (somatic markers) to potential outcomes. This idea, when placed in the context of interoceptive inference, corresponds to the guidance of behavioural (allostatic) responses towards the resolution of interoceptive prediction error (Gu & Fitzgerald 2014; Seth 2014a). It follows that intuitive decisions should be affected by the degree to which an individual maintains accurate predictive models of his or her own interoceptive states; see Dunn et al. 2010, Sokol-Hessner et al. 2014 for evidence along these lines.

There are also important implications for disorders of emotion, selfhood, and decision-making. For example, anxiety may result from the chronic persistence of interoceptive prediction errors that resist top-down suppression (Paulus & Stein 2006). Dissociative disorders like alexithymia (the inability to describe one’s own emotions), and depersonalization and derealisation (the loss of sense of reality of the self and world) may also result from dysfunctional interoceptive inference, perhaps manifest in abnormally low interoceptive precision expectations (Seth 2013; Seth et al. 2011). In terms of decision-making, it may be productive to think of addiction as resulting from dysfunctional active inference, whereby strong interoceptive priors are confirmed through action, overriding higher-order or hyper-priors relating to homeostasis and organismic integrity. It has even been suggested that

autism spectrum disorders may originate in aberrant encoding of the salience or precision of interoceptive prediction errors (Quattrocki & Friston 2014). The reasoning here is that aberrant salience during development could disrupt the assimilation of interoceptive and exteroceptive cues within generative models of the “self”, which would impair a child’s ability to properly assign salience to socially relevant signals.

3.3 The predictive embodied self

The maintenance of physiological homeostasis solely through direct autonomic regulation is obviously limited: behavioural (allostatic) interactions with the world are necessary if the organism is to avoid surprising physiological states in the long run. The ability to deploy adaptive behavioural responses mandates the original Helmholtzian view of perception-as-inference, which has been the primary setting for the development of PP so far. A critical but arguably overlooked middle ground, which mediates between physiological state variables and the external environment, is the *body*. On one hand, the body is the material vehicle through which behaviour is expressed, permitting allostatic interactions to take place. On the other, the body is itself an essential part of the organismic system, the homeostatic integrity of which must be maintained. In addition, the experience of owning and identifying with a particular body is a key component of being a conscious self (Apps & Tsakiris 2014; Blanke & Metzinger 2009; Craig 2009; Limanowski & Blankenburg 2013; Seth 2013).

It is tempting to ask whether common predictive mechanisms could underlie not only classical exteroceptive perception (like vision) and interoception (see above), but also their integration in supporting conscious and unconscious representations of the body and self (Seth 2013). The significance of this question is underlined by realising that just as the brain has no direct access to causal structures in the external environment, it also lacks direct access to its own body. That is, given that the brain is in the business of inferring the causal sources of

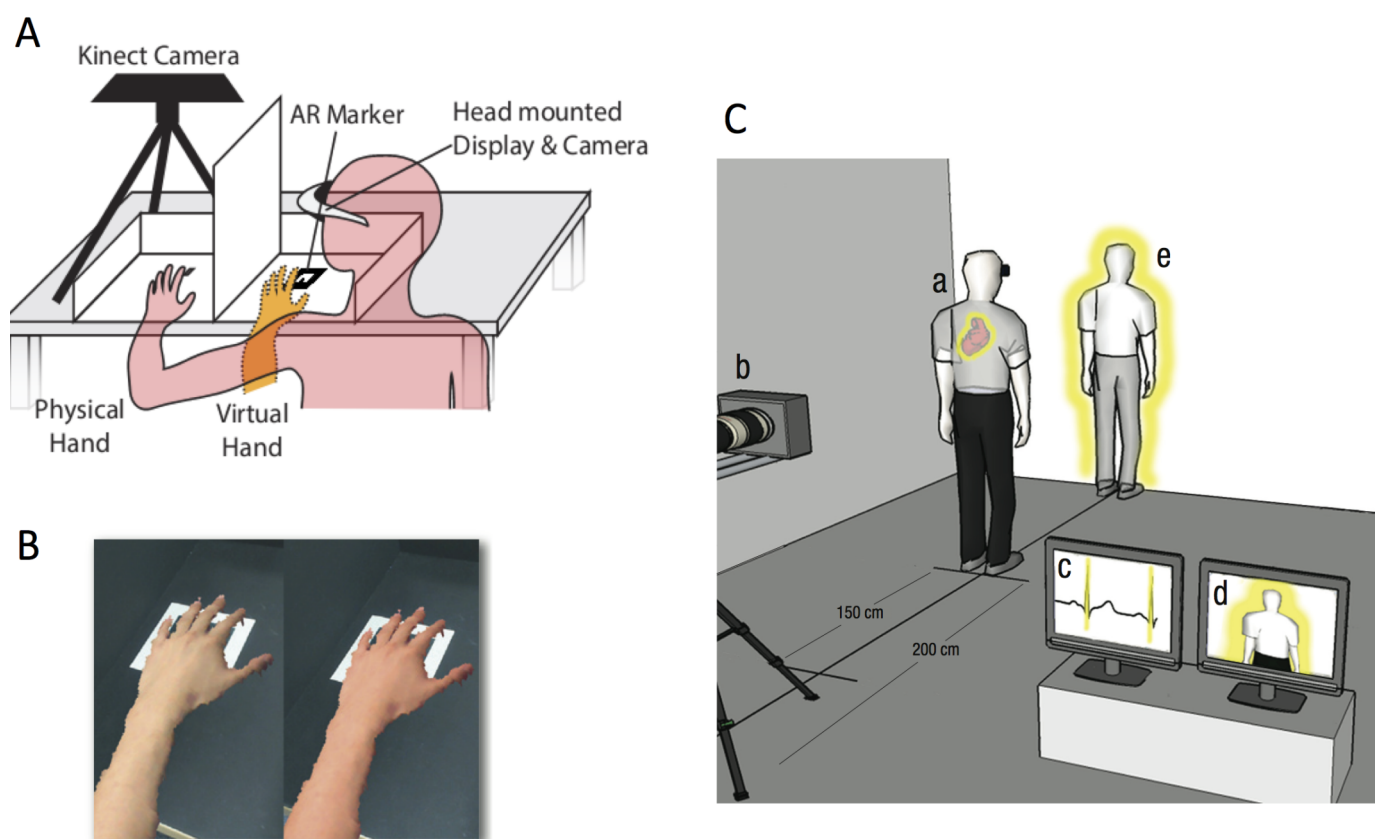


Figure 5: The interaction of interoceptive and exteroceptive signals in shaping the experience of body ownership. **A.** Set-up for applying cardio-visual feedback in the rubber hand illusion. A Microsoft Kinect obtains a real-time 3D model of a subject's left hand. This is re-projected into the subject's visual field using a head-mounted display and augmented reality (AR) software. **B.** The colour of the virtual hand is modulated by the subject's heart-beat. **C.** A similar set-up for the full-body illusion whereby a visual image of a subject's body is surrounded by a halo pulsing either in time or out of time with the heartbeat. Panels A and B are adapted from [Suzuki et al. \(2013\)](#); panel C is adapted from [Aspell et al. \(2013\)](#).

sensory signals, a key challenge emerges when distinguishing those signals that pertain to the body from those that originate from the external environment. A clue to how this challenge is met is that the physical body, unlike the external environment, constantly generates and receives internal input via its interoceptive and proprioceptive systems ([Limanowski & Blankenburg 2013](#); [Metzinger 2003](#)). This suggests that the experienced body (and self) depends on the brain's best guess of the causes of those sensory signals most likely to be “me” ([Apps & Tsakiris 2014](#)), across interoceptive, proprioceptive, and exteroceptive domains ([Figure 4](#)).

There is now considerable evidence that the *experience of body ownership* is highly plastic and depends on the multisensory integration of body-related signals ([Apps &](#)

[Tsakiris 2014](#); [Blanke & Metzinger 2009](#)). One classic example is the *rubber hand illusion*, where the stroking of an artificial hand synchronously with a participant's real hand, while visual attention is focused on the artificial hand, leads to the experience that the artificial hand is somehow part of the body ([Botvinick & Cohen 1998](#)). According to current multisensory integration models, this change in the experience of body ownership is due to correlation between vision and touch overriding conflicting proprioceptive inputs ([Makin et al. 2008](#)). Through the lens of PP, this implies that prediction errors induced by multisensory conflicts will over time update self-related priors ([Apps & Tsakiris 2014](#)), with different signal sources (vision, touch, proprioception) each precision-weighted according to their expected reliability, and all in

the setting of strong prior expectations for correlated input.⁵

While the potential for exteroceptive multisensory integration to modulate the experience of body ownership has been extensively explored both for the ownership of body parts and for the experience of ownership of the body as a whole (for reviews, see Apps & Tsakiris 2014; Blanke & Metzinger 2009), only recently has attention been paid to interactions between interoceptive and exteroceptive signals. Initial evidence in this line of investigation was indirect, for example showing correlation between susceptibility to the rubber hand illusion and individual differences in the ability to perceive interoceptive signals (“interoceptive sensitivity”, typically indexed by heartbeat detection tasks; Tsakiris et al. 2011). Other relevant studies have shown that body ownership illusions lead to temperature reductions in the corresponding body parts, perhaps reflecting altered active autonomic inference (Moseley et al. 2008; Salomon et al. 2013).

Emerging evidence now points more directly towards the predictive multisensory integration of interoceptive and exteroceptive signals in shaping the experience of body ownership. Two recent studies have taken advantage of so-called “cardio-visual synchrony” where virtual-reality representations of body parts (Suzuki et al. 2013) or the whole body (Aspell et al. 2013) are modulated by simultaneously recorded heartbeat signals, with the modulation either in-time or out-of-time with the actual heartbeat (Figure 5). These data suggest that statistical correlations between interoceptive (e.g., cardiac) and exteroceptive (e.g., visual) signals can lead to the updating of predictive models of self-related signals through (hierarchical) minimization of prediction error, just as happens for purely exteroceptive multisensory conflicts in the classic rubber hand illusion.

While these studies underline the plausibility of common predictive mechanisms underlying emotion, selfhood, and perception, many open questions nevertheless remain. A key challenge is to detail the underlying neural opera-

tions. Though a detailed analysis is beyond the scope of the present paper, it is worth noting that attention is increasingly focused on the insular cortex (especially its anterior parts) as a potential source of interoceptive predictions, and also as a comparator registering interoceptive prediction errors. The anterior insula has long been considered a major cortical locus for the integration of interoceptive and exteroceptive signals (Craig 2003; Singer et al. 2009); it is strongly implicated in interoceptive sensitivity (Critchley et al. 2004); it is sensitive to interoceptive prediction errors—at least in some contexts (Paulus & Stein 2006); and it has a high density of so-called “von Economo” neurons,⁶ which have been frequently though circumstantially associated with consciousness and selfhood (Critchley & Seth 2012; Evrard et al. 2012).

3.4 Active inference, self-modeling, and evolutionary robotics

What role might *active* inference play in predictive self-modelling? Autonomic changes during illusions of body ownership (see above) are consistent with active inference; however they do not speak directly to its function. In the classic rubber hand illusion, hand or finger movements can be considered active inferential tests of self-related hypotheses. If these movements are not reflected in the “rubber hand”, the illusion is destroyed—presumably because predicted visual signals are not confirmed (Apps & Tsakiris 2014). However, if hand movements are mapped to a virtual “rubber hand”—through clever use of virtual and augmented reality—the illusion is in fact strengthened, presumably because the multisensory correlation of peri-hand visual and proprioceptive signals constitutes a more stringent test of the perceptual hypothesis of ownership of the virtual hand (Suzuki et al. 2013). This introduces the idea that active inference is not simply about confirming sensory predictions but also involves seeking “disruptive” actions that are most informative with respect to testing current predictions,

⁵ Interestingly the expectation of perceptual correlations seems to be sufficient for inducing the rubber hand illusion (Ferri et al. 2013).

⁶ These are long-range projection neurons found selectively in hominid primates and certain other species.

and/or at disambiguating competing predictions (Gregory 1980). A nice example of how this happens in practice comes from *evolutionary robotics*⁷—which is obviously a very different literature, though one that inherits directly from the cybernetic tradition.

In a seminal 2006 study, Josh Bongard and colleagues described a four-legged “starfish” robot that engaged in a process much like active inference in order to model its own morphology so as to be able to control its movement and attain simple behavioural goals (Bongard et al. 2006). While there are important differences between evolutionary robotics and (active) Bayesian inference, there are also broad similarities; importantly, both can be cast in terms of model selection and optimization.

The basic cycle of events is shown in Figure 6. The robot itself is shown in the centre (A). The goal is to develop a controller capable of generating forward movement. The challenge is that the robot’s morphology is unknown to the robot itself. The system starts with a range of (generic prior) potential self-models (B), here specified by various configurations of three-dimensional physics engines. The robot performs a series of initially random actions and evaluates its candidate self-models on their ability to predict the resulting proprioceptive afferent signals. Even though all initial models will be wrong, some may be better than others. The key step comes next. The robot evaluates new candidate actions *on the extent to which the current best self-models make different predictions as to their (proprioceptive) consequences*. These disambiguating actions are then performed, leading to a new ranking of self-models based on their success at proprioceptive prediction. This ranking, via the evolutionary robotics methods of mutation and replication, gives rise to a new population of candidate self-models. The upshot is that the system swiftly develops accurate self-models that can be used to generate controllers enabling movement (D). An interesting feature of this process is that it is

highly resilient to unexpected perturbations. For instance, if a leg is removed then proprioceptive prediction errors will immediately ensue. As a result, the system will engage in another round of self-model evolution (including the co-specification of competing self-models and disambiguating actions) until a new, accurate, self-model is regained. This revised self-model can then be used to develop a new gait, allowing movement, even given the disrupted body (E, F).⁸

This study emphasizes that the operational criterion for a successful self-model is not so much its fidelity to the physical robot, but rather its ability to predict sensory inputs under a repertoire of actions. This underlines that predictive models are recruited for the control of behaviour (as cybernetics assumes) and not to furnish general-purpose representations of the world or the body.

The study also provides a concrete example of how actions can be performed, not to achieve some externally specified goal, but to permit inference about the system’s own physical instantiation. Bayesian or not, this implies active inference. Indeed, perhaps its most important contribution is that it highlights how active inference can prescribe *disruptive* or *disambiguating* actions that generate sensory prediction errors under competing hypotheses, and not just actions that seek to confirm sensory predictions. This recalls models of attention based on maximisation of Bayesian surprise (Itti & Baldi 2009), and is equivalent to hypothesis testing in science, where the best experiments are those concocted on the basis of being most likely to falsify a given hypothesis (disruptive) or distinguish between competing hypotheses (disambiguating). It also implies that agents encode predictions about the likely sensory consequences of a range of potential actions, allowing the selection of those actions likely to be the most disruptive or disambiguating. This concept of a *counterfactually-equipped predictive model* bring us nicely to our next topic: so-called *en-active* cognitive science and its relation to PP.

⁷ Evolutionary robotics involves the use of population-based search procedures (genetic algorithms) to automatically specify control architectures (and/or morphologies) of mobile robots. For an excellent introduction see (Bongard 2013).

⁸ Videos showing the evolution of both gait and self-model are available from http://creativemachines.cornell.edu/emergent_self_models

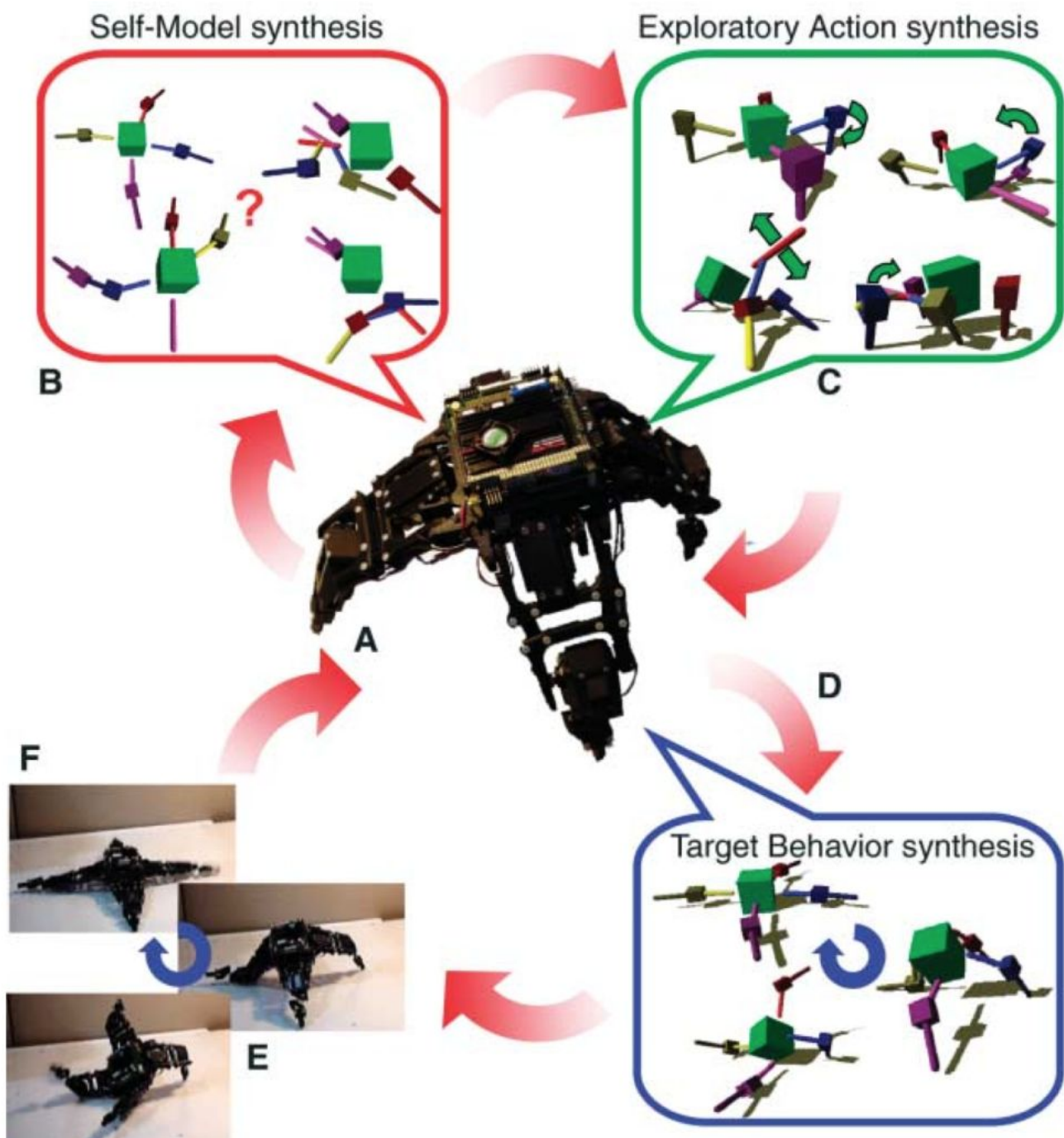


Figure 6: An evolutionary-robotics experiment demonstrating continuous self-modelling [Bongard et al. \(2006\)](#). See text for details. Reproduced with permission.

4 Predictive processing and enactive cognitive science

4.1 Enactive theories, weak and strong

The idea that the brain relies on internal representations or models of extra-cranial states of affairs has been treated with suspicion ever since the limitations of “good old fashioned arti-

ficial intelligence” became apparent ([Brooks 1991](#)). Many researchers of artificial intelligence have indeed returned to cybernetics as an alternative framework in which closely coupled feedback loops, leveraging invariants in brain-body-world interactions, obviate the need for detailed internal representations of external properties ([Pfeifer & Scheier 1999](#)). The evolutionary robotics methodology just described is

often coupled with simple dynamical neural networks in order to realize controllers that are tightly embodied and embedded in just this way (Beer 2003). Within cognitive science, such anti-representationalism is most vociferously defended by the movement variously known as “enactive” (Noë 2004), “embodied” (Gallese & Sinigaglia 2011), or “extended” (Clark & Chalmers 1998) cognitive science. Among these approaches, it is enactivism that is most explicitly anti-representationalist. While enactive theorists might agree that adaptive behaviour requires organisms and control structures that are systematically sensitive to statistical structures in their environment, most will deny that this sensitivity implies the existence and deployment of any “inner description” or model of these probabilistic patterns (Chemero 2009; Hutto & Myin 2013).

This tradition has weak and strong expressions. At the weak extreme is the truism that perception, cognition, and behaviour—and their underlying mechanisms—cannot be understood without a rich appreciation of the roles of the body, the environment, and the structured interactions that they support (Clark 1997; Varela et al. 1993). Weak enactivism is eminently compatible with PP, as seen especially with emerging versions of PP that stress embodiment through self-modelling and interoception, and which emphasize the importance of agent-environment coupling (embeddedness) through active inference. At the other extreme lie claims that explanations based on internal representations or models of any sort are fundamentally misguided, and that a new explicitly non-representational vocabulary is needed in order to make sense of the relations between brains, bodies, and the world (O’Regan et al. 2005). Strong enactivism is by definition incompatible with PP since it rejects the core concept of the internal model.

4.2 Sensorimotor contingency theory

A landmark in the strongly enactive approach is SMC (sensorimotor contingency) theory, which says that perception depends on the “practical mastery” of sensorimotor dependencies relevant

to behaviour (O’Regan & Noë 2001). In brief, SMC theory claims that experience and perception are not things that are “generated” by the brain (or by anything else for that matter) but are, rather, “skills” consisting of fluid patterns of on-going interaction with the environment (O’Regan & Noë 2001). For instance, on SMC theory the conscious visual experience of redness is given by *the exercise of practical mastery of the laws governing how interactions with red things unfold* (these laws being the “SMC”s). The theory is not, however, limited to vision: the experiential quality of the softness of a sponge would be given by (practical mastery of) the laws governing its squishiness upon being pressed.

Two aspects of SMC theory deserve emphasis here. The first is that the concept of an SMC rightly underlines the close coupling of perception and action and the critical importance of ongoing agent-environment interaction in structuring perception, action, and behaviour. This is inherited from Gibsonian notions of perceptual affordance (Gibson 1979) and has certainly advanced our understanding of why different kinds of perceptual experience (vision, smell, touch, etc.) have different qualitative characters.

The second is that *mastery* of an SMC requires an essentially *counterfactual* knowledge of relations between particular actions and the resulting sensations. In vision, for instance, mastery entails an implicit knowledge of the ways in which moving our eyes and bodies would reveal additional sensory information about perceptual objects (O’Regan & Noë 2001). Here SMC theory has made an important contribution to our understanding of *perceptual presence*. Perceptual presence refers to the property whereby (in normal circumstances) perceptual contents appear as subjectively real, that is, as *existing*. For example, when viewing a tomato, we see it as real inasmuch as we seem to be perceptually aware of some of its parts (e.g., its back) that are not currently causally impacting our sensory surfaces. Looking at a picture of a tomato does not give rise to the same subjective impression of realness. But how can we be aware of parts of the tomato that, strictly speaking, we do not

see? SMC theory says the answer lies in our (implicit) mastery of SMCs, which relate potential actions to their likely sensory effects; and it is in this sense that we can be perceptually aware of parts of the tomato that we cannot actually see (Noë 2006).

SMC theory has often been set against naïve representationalist theories in cognitive science that propose such things as “pictures in the head” or that (like good-old-fashioned-AI) treat accurate representations of external properties as general-purpose goal states for cognition. This is all to the good. Yet by dispensing with implementation-level concepts such as predictive inference, it struggles with the important question of what exactly is going on in our heads during the exercise of mastery of a sensorimotor contingency.⁹

4.3 Predictive perception of sensorimotor contingencies

A powerful response is given by integrating SMC theory with PP, in the guise of PPSMC (Predictive Perception of SensoriMotor Contingencies; Seth 2014b). An extensive development of PPSMC is given elsewhere (see Seth 2014b plus commentaries and response). Here I summarize the main points. First, recall that under PP prediction errors can be minimized either by updating perceptual predictions or by performing actions, where actions are generated through the resolution of proprioceptive prediction errors. Also recall that PP is inherently hierarchical, so that at some hierarchical level predictive models will encode multimodal and even amodal expectations linking exteroceptive (sensory) and proprioceptive (motor) sensations. These models generate predictions about linked sequences of sensory and proprioceptive (and possibly interoceptive) inputs corresponding to specific actions, with predictions becoming increasingly modality-specific at lower hierarchical levels. These multi-level predictive models can

therefore be understood as instantiating the implicit sub-personal knowledge of sensorimotor constructs underlying SMCs and their acquisition. Put simply, hierarchical active inference implies the existence of predictive models encoding information very much like that required by SMC theory.

The next step is to incorporate the notion of *mastery* of SMCs, which, as mentioned, implies an essentially counterfactual kind of implicit knowledge. The simple solution is to augment the predictive models that animate PP with counterfactual probability densities.¹⁰ As introduced earlier (section 4.1), counterfactually-equipped predictive models encode not only the likely causes of current sensory input, but also the likely causes of fictive sensory inputs conditioned on possible but not executed actions. That is, they encode how sensory inputs (and their expected precisions) would change on the basis of a repertoire of possible actions (expressed as proprioceptive predictions), even if those actions are not performed. The counterfactual encoding of expected precision is important here, since it is on this basis that actions can be selected for their likelihood of minimizing the conditional uncertainty associated with a perceptual prediction. There is a mathematical basis for manipulating counterfactual beliefs of this kind, as shown in a recent model where counterfactual PP drives oculomotor control during visual search (Friston 2014; Friston et al. 2012).¹¹ Here the main point is that counterfactually-rich predictive models supply just what is needed by SMC theory: an answer to the question of what is going on inside our heads during the exercise of mastery of SMCs.

Counterfactual PP makes sense from several perspectives (Seth 2014b). As mentioned above, it provides a neurocognitive operationalisation of the notion of mastery of SMCs that is central to enactive cognitive science. In doing so it dissolves apparent tensions between enactive

⁹ At a recent symposium of the AISB society that focused on SMC theory, it was stated that “the main question is how to get the brain into view from an enactive/sensorimotor perspective. [...] Addressing this question is urgently needed, for there seem to be no accepted alternatives to representational interpretations of the inner processes” (O’Regan & Dagenaar 2014).

¹⁰ See Beaton (2013) for a distinct approach to incorporating counterfactual ideas in SMC theory. Beaton’s approach remains squarely within the strongly enactivist tradition.

¹¹ There are also some challenges lying in wait here. For instance, it is not immediately clear how important assumptions like the Laplace approximation can generalize to the multimodal probability distributions entailed by counterfactual PP (Otworowska et al. 2014).

cognitive science and approaches grounded in the Bayesian brain, but only at the price of rejecting the strong enactivist's insistence that internal models or representations—of any sort—are unacceptable.¹² PPSMC also provides a solution to the challenge of accounting for perceptual presence within PP. The idea here is that perceptual presence corresponds to the *counterfactual richness* of predictive models. That is, perceptual contents enjoy presence to the extent that the corresponding predictive models encode a rich repertoire of counterfactual relations linking potential actions to their likely sensory consequences.¹³ In other words, we experience normal perception as world-revealing precisely because the predictive models underlying perceptual content specify a rich repertoire of counterfactually explicit probability densities encoding the mastery of SMCs.

A good test of PPSMC is whether it can account for cases where normal perceptual presence is lacking. An important example is synaesthesia, where it is widely reported that synaesthetic “concurrents” (e.g., the inexistent colours sometimes perceived along with achromatic grapheme inducers) are not experienced as being part of the world (i.e., synaesthetes generally retain intact reality testing with respect to their concurrent experiences). PPSMC explains this by noticing that predictive models related to synaesthetic concurrents are counterfactually *poor*. The hidden (environmental) causes giving rise to concurrent-related sensory signals do not embed a rich and deep statistical structure for the brain to learn. In particular, there is very little sense in which synaesthetic concurrents depend on active sampling of their hidden causes. According to PPSMC, it is this comparative *counterfactual poverty* that explains why synaesthetic concurrents lack perceptual presence. SMC theory itself struggles to account for this phenomenon—not least because it struggles to account for synaesthesia in the first place (Gray 2003).

¹² There is a more dramatic conflict with “radical” versions of enactivism, in which mental processes, and in some cases even their material substrates, are allowed to extend beyond the confines of the skull (Hutto & Myin 2013).

¹³ Presence may also depend on the hierarchical depth of predictive models inasmuch as this reflects object-related invariances in perception. For further discussion see commentaries and response to (Seth 2014b).

There are some challenges to thinking that perceptual presence uniquely depends on counterfactual richness. One might think that the more familiar one is with an object, the richer the repertoire of counterfactual relations that will be encoded. If so, the more familiar one is with an object, the more it should appear to be real. But *prima facie* it is not clear that familiarity and perceptual presence go hand-in-hand like this.¹⁴ Also, some perceptual experiences (like the experience of a blue sky) can seem highly perceptually present despite engaging an apparently poor repertoire of counterfactual relations linking sensory signals to possible actions. An initial response is to consider that presence might depend not on counterfactual richness *per se*, but on a “normalized” richness based on higher-order expectations of counterfactual richness (which would be low for the blue sky, for instance). These considerations also point to potentially important distinctions between perceived *objecthood* and perceived *presence*, a proper treatment of which moves beyond the scope of the present paper.

5 Active inference

5.1 Counterfactual PP and active inference

Active inference has appeared repeatedly as an important concept throughout this paper. Yet it is more difficult to grasp than the basics of PP, which involve passive predictive inference. This is partly because several senses of active inference can be distinguished, which have not previously been fully elaborated.

In general, active inference can be harnessed to drive action, or to improve perceptual predictions. In the former case, actions emerge from the minimization of proprioceptive prediction errors through engaging classical reflex arcs (Friston et al. 2010). This implies the existence of generative models that predict time-varying flows of proprioceptive inputs (rather than just end-points), and also the transient reduction of expected precision of proprioceptive prediction

¹⁴ Thanks to my reviewers for raising this provocative point.

errors, corresponding to sensory attenuation (Brown et al. 2013).

In the latter case, actions are engaged in order to generate new sensory samples, with the aim of minimizing uncertainty in perceptual predictions. This can be achieved in several different ways, as is apparent by analogy with experimental design in scientific hypothesis testing. Actions can be selected that (i) are expected to *confirm* current perceptual hypotheses (Friston et al. 2012); (ii) are expected to *disconfirm* such hypotheses; or (iii) are expected to *disambiguate* between competing hypotheses (Bongard et al. 2006). A scientist may perform different experiments when attempting to find evidence against a current hypothesis than when trying to decide between different hypotheses. In just the same way, active inference may prescribe different sampling actions for these different objectives.

These distinctions underline that active inference *implies* counterfactual PP. In order for a brain to select those actions most likely to confirm, disconfirm, or decide between current predictive model(s), it is necessary to encode expected sensory inputs and precisions related to potential (but not executed) actions. This is evident in the example of oculomotor control described earlier (Friston et al. 2012). Here, saccades are guided on the basis of the expected precision of sensory prediction errors so as to minimize the uncertainty in current perceptual predictions. Note that this study retained the higher-order prior that only a single perceptual prediction exists at any one time, precluding active inference in its disambiguatory sense.

Several related ideas arise in connection with these new readings of active inference. Seeking disconfirmatory or disruptive evidence is closely related to maximizing Bayesian surprise (Itti & Baldi 2009). This also reminds us that the best statistical models are usually those that successfully account for the most variance with the fewest degrees of freedom (model parameters), not just those that result in low residual error *per se*. In addition, disambiguating competing hypotheses moves from Bayesian model selection and optimization to model comparison, where arbitration among

competing models is mediated by trade-offs between accuracy and model complexity (Rosa et al. 2012).

The information-seeking (or “infotropic”¹⁵) role of active inference puts a different gloss on the free energy principle, which had been interpreted simply as minimization of prediction error. Rather, now the idea is that systems best ensure their long-run survival by inducing the *most predictive* model of the causes of sensory signals, and this requires disruptive and/or disambiguating active inference, in order to always put the current-best model to the test. This view helps dissolve worries about the so-called “dark room problem” (Friston et al. 2012), in which prediction error is minimized by predicting something simple (e.g., the absence of visual input) and then trivially confirming this prediction (e.g., by closing one’s eyes).¹⁶ Previous responses to this challenge have appealed to the idea of higher-order priors that are incompatible with trivial minimization of lower-level prediction errors: closing one’s eyes (or staying put in a dark room) is not expected to lead to homeostatic integrity on average and over time (Friston et al. 2012; Hohwy 2013). It is perhaps more elegant to consider that disruptive and disambiguatory active inferences imply exploratory sampling actions, independent of any higher-order priors about the dynamics of sensory signals *per se*. Further work is needed to see how cost functions reflecting infotropic active inference can be explicitly incorporated into PP and the free energy principle.

5.2 Active interoceptive inference and counterfactual PP

What can be said about counterfactual PP and active inference when applied to *interoception*? Is there a sense in which predictive models underlying emotion and mood encode counterfactual associations linking fictive interoceptive signals (and their likely causes) to autonomic or allostatic controls? And if so, what phenomeno-

¹⁵ Chris Thornton came up with this term (personal communication).

¹⁶ The term “dark room problem” comes from the idea that a free-energy-minimizing (or surprise-avoiding) agent could minimize prediction error just by finding an environment that lacks sensory stimulation (a “dark room”) and staying there.

logical dimensions of affective experience depend on these associations? While these remain open questions, we can at least sketch the territory.

We have seen that active inference in exteroception *implies* counterfactual processing, so that actions can be chosen according to their predicted effects in terms of (dis)confirming or disambiguating sensory predictions. The same argument applies to interoception. For active interoceptive inference to effectively disambiguate predictive models, or (dis)confirm interoceptive predictions, predictive models must be equipped with counterfactual associations relating to the likely effects of autonomic or (at higher hierarchical levels) allostatic controls. At least in this sense, interoceptive inference then also involves counterfactual expectations.

That said, there are likely to be substantial differences in how counterfactual active inference plays out in interoceptive settings. For instance, it may not be adaptive (in the long run) for organisms to continually attempt to disconfirm current interoceptive predictions, assuming these are compatible with homeostatic integrity. To put it colloquially, we do not want to drive our essential variables continually close to viability limits, just to check whether they are always capable of returning. This recalls our earlier point (section 4.1) that predictive control is more naturally applicable to interoception than exteroception, given the imperative of maintaining the homeostasis of essential variables. In addition, the causal structure of counterfactual associations encoded by interoceptive predictive models is undoubtedly very different than in cases like vision. These differences may speak to the substantial phenomenological differences in the kind of perceptual presence associated with these distinct conscious contents (Seth et al. 2011).

6 Conclusion

This paper has surveyed predictive processing (PP) from the unusual viewpoint of cybernetic origins in active homeostatic control (Ashby 1952; Conant & Ashby 1970). This shifts the perspective from perceptual inference as fur-

nishing representations of the external world for the consumption of general-purpose cognitive mechanisms, towards model-based predictive control as a primary survival imperative from which perception, action, and cognition ensue. This view is aligned with the free energy principle (Friston 2010); however it attempts to account for specific cognitive and phenomenological properties, rather than for adaptive systems in general. Several implications follow from these considerations. Emotion becomes a process of active interoceptive inference (Seth 2013)—a process that also recruits autonomic regulation and influences intuitive decision-making through behavioural allostasis. A common predictive principle underlying interoception and exteroception also provides an integrative view of the neurocognitive mechanisms underlying embodied selfhood, in particular the experience of body ownership (Apps & Tsakiris 2014; Limanowski & Blankenburg 2013; Suzuki et al. 2013). In this view, the experience of embodied selfhood is specified by the brain’s “best guess” of those signals most likely to be “me” across exteroceptive and interoceptive domains. From the perspective of cybernetics the embodied self is both that which needs to be homeostatically maintained and also the medium through which allostatic interactions are expressed.

A second influential line deriving from cybernetics sets PP within the broader context of model-based versus enactivist perspectives on cognitive science. On one hand, cybernetics has been cited in support of non-representational cognitive science in virtue of its showing how simple mechanisms can give rise to complex and apparently goal-directed behaviour by capitalizing on agent-environment interactions, mediated by the body (Pfeifer & Scheier 1999). On the other, the cybernetic legacy shows how PP can put mechanistic flesh on the philosophical bones of enactivism, but only by embracing a finessed form of representationalism (Seth 2014b). A key concept within enactive cognitive science is that of mastery of sensorimotor contingencies (SMCs). This concept is useful for understanding the qualitative character of distinct perceptual modalities, yet as expressed within enactivism it lacks a firm implementation basis. “Pre-

dictive Perception of SensoriMotor Contingencies” (PPSMC) addresses this challenge by proposing that SMCs are implemented by predictive models of sensorimotor relations, underpinned by the continuity between perception and action entailed by active inference. *Mastery* of sensorimotor contingencies depends on predictive models of counterfactual probability densities that specify the likely causes of sensory signals that *would* occur *were* specific actions taken. By relating PP to key concepts in enactivism, this theory is able to account for phenomenological features well treated by the latter, such as the experience of perceptual presence (and its absence in cases like synaesthesia).

Considering these issues leads to distinct readings of active inference, which at its most general implies the selective sampling of sensory signals to minimize uncertainty about perceptual predictions. At a finer grain, active inference can involve performing actions to confirm current predictions, to disconfirm current predictions, or to disambiguate competing predictions. These different senses rest on the concept of counterfactually-equipped predictive models; and they generalize the free energy principle to include Bayesian-model comparison as well as optimization and inference.

In summary, the ideas outlined in this paper provide a distinctive integration of predictive processing, cybernetics, and enactivism. This rich blend dissolves apparent tensions between internalist and enactivist (model-based and model-free) views on the neural mechanisms underlying perception, cognition, and action; it elaborates common predictive mechanisms underlying perception and control of self and world; it provides a new view of emotion as active interoceptive inference, and it shows how “counterfactual” predictive processing can account for the phenomenology of conscious presence and its absence in specific situations. It also finesses the concept of active inference to engage distinct forms of hypothesis testing that prescribe different sampling actions (one bonus is that the “dark room problem” is elegantly solved). At the same time, new and difficult challenges arise in validating these ideas experi-

mentally and in distinguishing them from alternative explanations that do not rely on internally-realised inferential mechanisms.

Acknowledgements

I am grateful to the Dr. Mortimer and Theresa Sackler Foundation, which supports the work of the Sackler Centre for Consciousness Science. This work was also supported by ERC FP7 grant CEEDs (FP7-ICT-2009-5, 258749). Many thanks to Thomas Metzinger and Jennifer Windt for inviting me to make this contribution, and for the insightful and helpful reviewer comments they solicited. I’m also grateful to Kevin O’Regan and Jan Dagenaar for inviting me to speak at a symposium entitled “Consciousness without inner models?” (London, April 2014), which provided a feisty forum for debating some of the ideas presented here.

References

- Adams, R. A., Shipp, S. & Friston, K. J. (2013). Predictions not commands: Active inference in the motor system. *Brain Structure and Function*, 218 (3), 611-643. [10.1007/s00429-012-0475-5](https://doi.org/10.1007/s00429-012-0475-5)
- Apps, M. A. & Tsakiris, M. (2014). The free-energy self: A predictive coding account of self-recognition. *Neuroscience and Biobehavioral Reviews*, 41, 85-97. [10.1016/j.neubiorev.2013.01.029](https://doi.org/10.1016/j.neubiorev.2013.01.029)
- Ashby, W. R. (1952). *Design for a brain*. London, UK: Chapman and Hall.
- (1956). *An introduction to cybernetics*. London, UK: Chapman and Hall.
- Aspell, J. E., Heydrich, L., Marillier, G., Lavanchy, T., Herbelin, B. & Blanke, O. (2013). Turning the body and self inside out: Visualized heartbeats alter bodily self-consciousness and tactile perception. *Psychological Science*, 24 (12), 2445-2453. [10.1177/0956797613498395](https://doi.org/10.1177/0956797613498395)
- Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P. & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, 76 (4), 695-711. [10.1016/j.neuron.2012.10.038](https://doi.org/10.1016/j.neuron.2012.10.038)
- Beaton, M. (2013). Phenomenology and embodied action. *Constructivist Foundations*, 8 (3), 298-313.
- Beer, R. D. (2003). The dynamics of active categorical perception in an evolved model agent. *Adaptive Behavior*, 11 (4), 209-243. [10.1177/1059712303114001](https://doi.org/10.1177/1059712303114001)

- Blanke, O. & Metzinger, T. (2009). Full-body illusions and minimal phenomenal selfhood. *Trends in Cognitive Sciences*, 13 (1), 7-13. [10.1016/j.tics.2008.10.003](https://doi.org/10.1016/j.tics.2008.10.003)
- Bongard, J. (2013). Evolutionary robotics. *Communications of the ACM*, 56 (8), 74-85. [10.1145/2493883](https://doi.org/10.1145/2493883)
- Bongard, J., Zykov, V. & Lipson, H. (2006). Resilient machines through continuous self-modeling. *Science*, 314 (5802), 1118-1121. [10.1126/science.1133687](https://doi.org/10.1126/science.1133687)
- Botvinick, M. & Cohen, J. (1998). Rubber hands ‘feel’ touch that eyes see. *Nature*, 391 (6669), 756-756. [10.1038/35784](https://doi.org/10.1038/35784)
- Braitenberg, V. (1984). *Vehicles: Experiments in synthetic psychology*. Cambridge, MA: MIT Press.
- Brooks, R. A. (1991). Intelligence without reason. In J. Mylopoulos & R. Reiter (Eds.) *Proceedings of the 12th international joint conference on artificial intelligence - volume 1* (pp. 569-595). San Francisco, CA: Morgan Kaufmann Publishers.
- Brown, H., Adams, R. A., Parees, I., Edwards, M. & Friston, K. J. (2013). Active inference, sensory attenuation and illusions. *Cognitive Processing*, 14 (4), 411-427. [10.1007/s10339-013-0571-3](https://doi.org/10.1007/s10339-013-0571-3)
- Chemero, A. (2009). *Radical embodied cognitive science*. Cambridge, MA: MIT Press.
- Clark, A. (1997). *Being there. Putting brain, body, and world together again*. Cambridge, MA: MIT Press.
- (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavior and Brain Sciences*, 36 (3), 181-204. [10.1017/S0140525X12000477](https://doi.org/10.1017/S0140525X12000477)
- (2015). Embodied prediction. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1-21). Frankfurt a. M., GER: MIND Group.
- Clark, A. & Chalmers, D. J. (1998). The extended mind. *Analysis*, 58 (1), 7-19. [10.1093/analys/58.1.7](https://doi.org/10.1093/analys/58.1.7)
- Conant, R. & Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 1 (2), 89-97.
- Craig, A. D. (2003). Interoception: The sense of the physiological condition of the body. *Current Opinion in Neurobiology*, 13 (4), 500-505. [10.1016/S0959](https://doi.org/10.1016/S0959)
- (2009). How do you feel now? The anterior insula and human awareness. *Nature Reviews Neuroscience*, 10 (1), 59-70. [10.1038/nrn2555](https://doi.org/10.1038/nrn2555)
- Critchley, H. D., Wiens, S., Rotshtein, P., Ohman, A. & Dolan, R. J. (2004). Neural systems supporting interoceptive awareness. *Nature Neuroscience*, 7 (2), 189-195. [10.1038/nrn1176](https://doi.org/10.1038/nrn1176)
- Critchley, H. D. & Seth, A. K. (2012). Will studies of macaque insula reveal the neural mechanisms of self-awareness? *Neuron*, 74 (3), 423-426. [10.1016/j.neuron.2012.04.012](https://doi.org/10.1016/j.neuron.2012.04.012)
- Damasio, A. (1994). *Descartes’ error*. London, UK: Mac Millan.
- Dunn, B. D., Galton, H. C., Morgan, R., Evans, D., Oliver, C., Meyer, M. & Dalgleish, T. (2010). Listening to your heart. How interoception shapes emotion experience and intuitive decision making. *Psychological Science*, 21 (12), 1835-1844. [10.1177/0956797610389191](https://doi.org/10.1177/0956797610389191)
- Dupuy, J.-P. (2009). *On the origins of cognitive science: The mechanization of mind*. Cambridge, MA: MIT Press.
- Evrard, H. C., Forro, T. & Logothetis, N. K. (2012). Von economo neurons in the anterior insula of the macaque monkey. *Neuron*, 74 (3), 482-489. [10.1016/j.neuron.2012.03.003](https://doi.org/10.1016/j.neuron.2012.03.003)
- Ferri, F., Chiarelli, A. M., Merla, A., Gallese, V. & Costantini, M. (2013). The body beyond the body: Expectation of a sensory event is enough to induce ownership over a fake hand. *Proceedings of the Royal Society B: Biological Sciences*, 280 (1765), 20131140-20131140. [10.1098/rspb.2013.1140](https://doi.org/10.1098/rspb.2013.1140)
- Fletcher, P. C. & Frith, C. D. (2009). Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, 10 (1), 48-58. [10.1038/nrn2536](https://doi.org/10.1038/nrn2536)
- Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360 (1456), 815-836. [10.1098/rstb.2005.1622](https://doi.org/10.1098/rstb.2005.1622)
- (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, 13 (7), 293-301. [10.1016/j.tics.2009.04.005](https://doi.org/10.1016/j.tics.2009.04.005)
- (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11 (2), 127-138. [10.1038/nrn2787](https://doi.org/10.1038/nrn2787)
- (2014). Active inference and agency. *Cognitive Neuroscience*, 5 (2), 119-121. [10.1080/17588928.2014.905517](https://doi.org/10.1080/17588928.2014.905517)
- Friston, K. J., Kilner, J. & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology - Paris*, 100 (1-3), 70-87. [10.1016/j.jphysparis.2006.10.001](https://doi.org/10.1016/j.jphysparis.2006.10.001)
- Friston, K. J., Daunizeau, J., Kilner, J. & Kiebel, S. J. (2010). Action and behavior: A free-energy formulation. *Biological Cybernetics*, 102 (3), 227-260. [10.1007/s00422-010-0364-z](https://doi.org/10.1007/s00422-010-0364-z)

- Friston, K. J., Adams, R. A., Perrinet, L. & Breakspear, M. (2012). Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology*, 3 (151), 1-20. [10.3389/fpsyg.2012.00151](https://doi.org/10.3389/fpsyg.2012.00151)
- Friston, K. J., Thornton, C. & Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in Psychology*, 3 (130), 1-7. [10.3389/fpsyg.2012.00130](https://doi.org/10.3389/fpsyg.2012.00130)
- Gallese, V. & Sinigaglia, C. (2011). What is so special about embodied simulation? *Trends in Cognitive Sciences*, 15 (11), 512-519. [10.1016/j.tics.2011.09.003](https://doi.org/10.1016/j.tics.2011.09.003)
- Gibson, J. J. (1979). *The ecological approach to visual perception*. Hillsdale, NJ: Lawrence Erlbaum.
- Gray, J. A. (2003). How are qualia coupled to functions? *Trends in Cognitive Sciences*, 7 (5), 192-194. [10.1016/S1364-6613\(03\)00077-9](https://doi.org/10.1016/S1364-6613(03)00077-9)
- Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 290 (1038), 181-197. [10.1098/rstb.1980.0090](https://doi.org/10.1098/rstb.1980.0090)
- Gu, X., Hof, P. R., Friston, K. J. & Fan, J. (2013). Anterior insular cortex and emotional awareness. *Journal of Comparative Neurology*, 521 (15), 3371-3388. [10.1002/cne.23368](https://doi.org/10.1002/cne.23368)
- Gu, X. & Fitzgerald, T. H. (2014). Interoceptive inference: Homeostasis and decision-making. *Trends in Cognitive Sciences*, 18 (6), 269-270. [10.1016/j.tics.2014.02.001](https://doi.org/10.1016/j.tics.2014.02.001)
- Hinton, G. E. & Dayan, P. (1996). Varieties of Helmholtz Machine. *Neural Networks*, 9 (8), 1385-1403. [10.1016/S0893](https://doi.org/10.1016/S0893)
- Hohwy, J. (2013). *The predictive mind*. Oxford, UK: Oxford University Press.
- (2015). The neural organ explains the mind. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1-22). Frankfurt a.M., GER: MIND Group.
- Hutto, D. & Myin, E. (2013). *Radicalizing enactivism: Basic minds without content*. Cambridge, MA: MIT Press.
- Itti, L. & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, 49 (10), 1295-1306. [10.1016/j.visres.2008.09.007](https://doi.org/10.1016/j.visres.2008.09.007)
- James, W. (1894). The physical basis of emotion. *Psychological Review*, 1, 516-529.
- Knill, D. C. & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27 (12), 712-719. [10.1016/j.tins.2004.10.007](https://doi.org/10.1016/j.tins.2004.10.007)
- Lee, T. S. & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A, Optics, image science and vision*, 20 (7), 1434-1448. [10.1364/JOSAA.20.001434](https://doi.org/10.1364/JOSAA.20.001434)
- Limanowski, J. & Blankenburg, F. (2013). Minimal self-models and the free energy principle. *Frontiers in Human Neurosciences*, 7 (547), 1-20. [10.3389/fnhum.2013.00547](https://doi.org/10.3389/fnhum.2013.00547)
- Makin, T. R., Holmes, N. P. & Ehrsson, H. H. (2008). On the other hand: Dummy hands and peripersonal space. *Behavioural Brain Research*, 191 (1), 1-10. [10.1016/j.bbr.2008.02.041](https://doi.org/10.1016/j.bbr.2008.02.041)
- Metzinger, T. (2003). *Being no one*. Cambridge, MA: MIT Press.
- Moseley, G. L., Olthof, N., Venema, A., Don, S., Wijers, M., Gallace, A. & Spence, C. (2008). Psychologically induced cooling of a specific body part caused by the illusory ownership of an artificial counterpart. *Proceedings of the National Academy of Sciences of the United States of America*, 105 (35), 13169-13173. [10.1073/pnas.0803768105](https://doi.org/10.1073/pnas.0803768105)
- Neal, R. M. & Hinton, G. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan (Ed.) *Learning in Graphical Models* (pp. 355-368). Dordrecht, NL: Kluwer Academic Publishers.
- Noë, A. (2004). *Action in perception*. Cambridge, MA: MIT Press.
- (2006). *Experience without the head*. Clarendon, NY: Oxford University Press.
- O'Regan, J. K. & Dagenaar, J. (2014). Consciousness without inner models: A sensorimotor account of what is going on in our heads. *Proceedings of the AISB*. <http://doc.gold.ac.uk/aisb50/>
- O'Regan, J. K., Myin, E. & Noë, A. (2005). Skill, corporality and alerting capacity in an account of sensory consciousness. *Progress in Brain Research*, 150, 55-68. [10.1016/S0079-6123\(05\)50005-0](https://doi.org/10.1016/S0079-6123(05)50005-0)
- O'Regan, J. K. & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24 (5), 939-1031.
- Otworowska, M., Kwisthout, J. & van Rooj, I. (2014). Counterfactual mathematics of counterfactual predictive models. *Frontiers in psychology: Consciousness Research*, 5 (801), 1-2. [10.3389/fpsyg.2014.00801](https://doi.org/10.3389/fpsyg.2014.00801)
- Paulus, M. P. & Stein, M. B. (2006). An insular view of anxiety. *Biological psychiatry*, 60 (4), 383-387. [10.1016/j.biopsych.2006.03.042](https://doi.org/10.1016/j.biopsych.2006.03.042)

- Pfeifer, R. & Scheier, C. (1999). *Understanding intelligence*. Cambridge, MA: MIT Press.
- Pickering, A. (2010). *The cybernetic brain: Sketches of another future*. Chicago, IL: University of Chicago Press.
- Ploghaus, A., Tracey, I., Gati, J. S., Clare, S., Menon, R. S., Matthews, P. M. & Rawlins, J. N. (1999). Dissociating pain from its anticipation in the human brain. *Science*, 284 (5422), 1979-1981. [10.1126/science.284.5422.1979](https://doi.org/10.1126/science.284.5422.1979)
- Pouget, A., Beck, J. M., Ma, W. J. & Latham, P. E. (2013). Probabilistic brains: Knowns and unknowns. *Nature Neuroscience*, 16 (9), 1170-1178. [10.1038/nm.3495](https://doi.org/10.1038/nm.3495)
- Quattrocki, E. & Friston, K. (2014). Autism, oxytocin and interoception. *Neuroscience and Biobehavioral Reviews*, 47C, 410-430. [10.1016/j.neubiorev.2014.09.012](https://doi.org/10.1016/j.neubiorev.2014.09.012)
- Rao, R. P. & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2 (1), 79-87. [10.1038/4580](https://doi.org/10.1038/4580)
- Rosa, M. J., Friston, K. J. & Penny, W. (2012). Post-hoc selection of dynamic causal models. *Journal of Neuroscience Methods*, 208 (1), 66-78. [10.1016/j.jneumeth.2012.04.013](https://doi.org/10.1016/j.jneumeth.2012.04.013)
- Salomon, R., Lim, M., Pfeiffer, C., Gassert, R. & Blanke, O. (2013). Full body illusion is associated with widespread skin temperature reduction. *Frontiers in Behavioral Neuroscience*, 7 (65), 1-11. [10.3389/fnbeh.2013.00065](https://doi.org/10.3389/fnbeh.2013.00065)
- Schachter, S. & Singer, J. E. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, 69, 379-399. [10.1037/h0046234](https://doi.org/10.1037/h0046234)
- Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, 17 (11), 565-573. [10.1016/j.tics.2013.09.007](https://doi.org/10.1016/j.tics.2013.09.007)
- (2014a). Interoceptive inference: From decision-making to organism integrity. *Trends in Cognitive Sciences*, 18 (6), 270-271. [10.1016/j.tics.2014.03.006](https://doi.org/10.1016/j.tics.2014.03.006)
- (2014b). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synaesthesia. *Cognitive Neuroscience*, 5 (2), 97-118. [10.1080/17588928.2013.877880](https://doi.org/10.1080/17588928.2013.877880)
- Seth, A. K. & Critchley, H. D. (2013). Interoceptive predictive coding: A new view of emotion? *Behavioral and Brain Sciences*, 36 (3), 227-228.
- Seth, A. K., Suzuki, K. & Critchley, H. D. (2011). An interoceptive predictive coding model of conscious presence. *Frontiers in Psychology*, 2 (395), 1-16. [10.3389/fpsyg.2011.00395](https://doi.org/10.3389/fpsyg.2011.00395)
- Singer, T., Critchley, H. D. & Preuschoff, K. (2009). A common role of insula in feelings, empathy and uncertainty. *Trends in Cognitive Sciences*, 13 (8), 334-340. [10.1016/j.tics.2009.05.001](https://doi.org/10.1016/j.tics.2009.05.001)
- Sokol-Hessner, P., Hartley, C. A., Hamilton, J. R. & Phelps, E. A. (2014). Interoceptive ability predicts aversion to losses. *Cognition and Emotion*, 1-7. [10.1080/02699931.2014.925426](https://doi.org/10.1080/02699931.2014.925426)
- Suzuki, K., Garfinkel, S. N., Critchley, H. D. & Seth, A. K. (2013). Multisensory integration across exteroceptive and interoceptive domains modulates self-experience in the rubber-hand illusion. *Neuropsychologia*, 51 (13), 2909-2917. [10.1016/j.neuropsychologia.2013.08.014](https://doi.org/10.1016/j.neuropsychologia.2013.08.014)
- Thompson, E. & Varela, F. J. (2001). Radical embodiment: Neural dynamics and consciousness. *Trends in Cognitive Sciences*, 5 (10), 418-425. [10.1016/S1364-6613\(00\)01750-2](https://doi.org/10.1016/S1364-6613(00)01750-2)
- Tsakiris, M., Tajadura-Jimenez, A. & Costantini, M. (2011). Just a heartbeat away from one's body: Interoceptive sensitivity predicts malleability of body-representations. *Proceedings. Biological sciences / The Royal Society*, 278 (1717), 2470-2476. [10.1098/rspb.2010.2547](https://doi.org/10.1098/rspb.2010.2547)
- Ueda, K., Okamoto, Y., Okada, G., Yamashita, H., Hori, T. & Yamawaki, S. (2003). Brain activity during expectancy of emotional stimuli: An fMRI study. *NeuroReport*, 14 (1), 51-55. [10.1097/01.wnr.0000050712.17082.1c](https://doi.org/10.1097/01.wnr.0000050712.17082.1c)
- Van Gelder, T. (1995). What might cognition be if not computation? *Journal of Philosophy*, 92 (7), 345-381. [10.2307/2941061](https://doi.org/10.2307/2941061)
- Varela, F., Thompson, E. & Rosch, E. (1993). *The embodied mind: Cognitive science and human experience*. Cambridge, MA: MIT Press.
- Verschure, P. F., Voegtlin, T. & Douglas, R. J. (2003). Environmentally mediated synergy between perception and behaviour in mobile robots. *Nature*, 425 (6958), 620-624. [10.1038/nature02024](https://doi.org/10.1038/nature02024)
- Wolpert, D. M. & Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nature Neuroscience*, 3 Suppl, 1212-1217. [10.1038/81497](https://doi.org/10.1038/81497)

Perceptual Presence in the Kuhnian-Popperian Bayesian Brain

A Commentary on Anil K. Seth

Wanja Wiese

Anil Seth's target paper connects the framework of PP (predictive processing) and the FEP (free-energy principle) to cybernetic principles. Exploiting an analogy to theory of science, Seth draws a distinction between three types of active inference. The first type involves confirmatory hypothesis-testing. The other types involve seeking disconfirming and disambiguating evidence, respectively. Furthermore, Seth applies PP to various fascinating phenomena, including perceptual presence. In this commentary, I explore how far we can take the analogy between explanation in perception and explanation in science.

In the first part, I draw a slightly broader analogy between PP and concepts in theory of science, by asking whether the Bayesian brain is Kuhnian or Popperian. While many aspects of PP are in line with Karl Popper's falsificationism, other aspects of PP conform to how Thomas Kuhn described scientific revolutions. Thus, there is both a sense in which the Bayesian brain is Kuhnian, and a sense in which it is Popperian. The upshot of these considerations is that falsification in PP can take many different forms. In particular, active inference can be used to falsify a model in more ways than identified by Seth.

In the second part of this commentary, I focus on Seth's PPSMCT (predictive processing account of sensorimotor contingency theory) and its application to perceptual presence, which assigns a crucial role to counterfactual richness. In my discussion, I question the significance of counterfactual richness for perceptual presence. First, I highlight an ambiguity inherent in Seth's descriptions of the target phenomenon (perceptual presence vs. objecthood). Then I suggest that counterfactual richness may not be the crucial underlying feature (of either perceptual presence or objecthood). Giving a series of examples, I argue that the degree of *represented causal integration* is an equally good candidate for accounting for perceptual presence (or objecthood), although more work needs to be done.

Keywords

Active inference | Binocular rivalry | Counterfactual richness | Cybernetics | Demarcation problem | Falsification | Free-energy principle | Naïve falsificationism | Objecthood | Paradigm change | Perceptual presence | Predictive processing | Rubber hand illusion | Scientific progress | Sensorimotor contingencies | Sophisticated falsificationism

1 Introduction

One of the relevant aspects of Seth's discussion is the way in which it highlights interesting links to theoretical precursors of PP. In doing so, he broadens the historical context in which the framework is usually situated. However, these considerations are not just relevant for the

history of science, they also constitute a theoretical underpinning of several ways in which Seth has recently developed PP accounts of various phenomena. Due to limited space, I can only address some of these here. In particular, I will focus on his three interpretations of active

Commentator

Wanja Wiese

wawiese@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Target Author

Anil K. Seth

a.k.seth@sussex.ac.uk

University of Sussex
Brighton, United Kingdom

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

inference, and on his PP account of perceptual presence. In so doing, I will also try to take the analogy between explanation in perception and explanation in science a little further than it has previously been taken.

In section 2, I will briefly summarize Seth's view on the connection between cybernetics and the free-energy principle. One of the results of his considerations is that a distinction can be drawn between three types of active inference. The first type involves confirmatory hypothesis-testing. The other types involve seeking disconfirming and disambiguating evidence, respectively. Seth does not say much about what it takes to disconfirm or falsify a hypothesis or model. Furthermore, he seems to suggest that not all types of active inference he distinguishes are currently part of PP (at least in the version described by Karl Friston's FEP): "[t]hese points represent significant developments of the basic infrastructure of PP" (Seth 2014, p. 3).¹ In section 3, I will provide clarification of the notion of falsification by referring to the works of Karl Popper, Imre Lakatos, and Thomas Kuhn. I will also provide examples to show that different types of falsification are part and parcel of PP, not extensions of the basic infrastructure. In section 4, I point out an ambiguity in Seth's account of perceptual presence (perceptual presence vs. objecthood). After this, I suggest that counterfactual richness may not be the crucial underlying feature (of either perceptual presence or objecthood). Giving a series of examples, I argue that the degree of *represented causal integration* is an equally good candidate for accounting for perceptual presence (or objecthood), although more work needs to be done.

2 Cybernetics and the free-energy principle

In his very rich target paper, Anil Seth calls attention to one of the less well-considered precursors of PP: cybernetics. A central concept of cybernetics is the notion of homeostasis, which denotes an equilibrium of the system's paramet-

ers. This equilibrium is maintained by keeping the system's essential variables, like levels of blood oxygenation or blood sugar (cf. Seth this collection, p. 7), within a certain range (cf. *ibid.* pp. 7-8.). The process of achieving homeostasis is called allostasis (cf. *ibid.* p. 8). Cybernetic systems are teleological, i.e., goal-directed, because they are always trying to reach and preserve homeostasis. This suggests that control is more important than perception (cf. *ibid.* p. 9), and, as Seth emphasizes, it prioritizes interoceptive control over exteroceptive control: the main goal is to control the system's essential variables; interaction with the world is only necessary to the extent that it affects these variables (*ibid.* pp. 9-10.).

The principles of cybernetics fit astonishingly well to ideas motivating Karl Friston's FEP (which can, in some respects, be seen as a generalization of predictive processing).² The fundamental assumption behind this principle is that biological systems seek to "maintain their states and form in the face of a constantly changing environment" (Friston 2010, p. 127). This is obviously similar to the goal of achieving homeostasis.³ Another focus of FEP is active inference, because action can reduce the surprisal of the agent's states (which is necessary to "resist a tendency to disorder", Friston 2009, p. 293); perceptual inference can only reduce the free-energy bound on surprise (Friston 2009, p. 294). This is in stark contrast with the Helmholtzian roots of PP, according to which action is primarily in the service of perception:

[...] wir beobachten unter fortdauernder eigener Thätigkeit, und gelangen dadurch zur Kenntniss des Bestehens eines gesetzlichen Verhältnisses zwischen unseren Innervationen und dem Präsentwerden der verschiedenen Eindrücke aus dem Kreise

¹ Unless stated otherwise, all page numbers refer to the target paper by Anil Seth.

² It is more general, because predictive processing only plays a role in it if combined with the Laplace approximation (which entails, roughly, that probability distributions are approximated by Gaussian distributions). This approximation, however, also turns FEP into a more specific version, by assuming that the brain codes probability distribution as Gaussian distributions (which is not entailed by the general predictive processing framework discussed in Clark 2013, for instance).

³ In fact, the free-energy principle seems to be partly inspired by cybernetic ideas. Friston (2010, p. 127), for instance, cites Ashby (1947) when explaining the motivation for FEP.

der zeitweiligen Präsentabilien. Jede unserer willkürlichen Bewegungen, durch die wir die Erscheinungsweise der Objecte abändern, ist als ein Experiment zu betrachten, durch welches wir prüfen, ob wir das gesetzliche Verhalten der vorliegenden Erscheinung, d.h. ihr vorausgesetztes Bestehen in bestimmter Raumordnung, richtig aufgefasst haben.⁴ (Helmholtz 1959, p. 39)

According to this view, the main target of action is to find confirmatory evidence for internally-generated hypotheses. In short, the contrast between these two views can be described as “action as hypothesis-testing” versus “action as predictive control”. Whereas the first seems to fit best to the Helmholtzian roots of PP (and puts action in the service of perception), the second seems to fit better to its cybernetic origins. Most notably, the free-energy principle combines both aspects, but assigns a pivotal role to action (perceptual inference only makes the free-energy bound on surprise tight, active inference leads to a further reduction of free energy, reducing surprise implicitly).

Seth compares model selection and optimization in evolutionary robotics to how these processes are implemented in active inference (pp. 14-15.). He cites the famous starfish robot developed by Josh Bongard, Victor Zykov, & Hod Lipson (2006) as an example. In a first phase, the robot generates multiple competing models of its own morphology and performs actions for which these models predict different sensory feedback. By comparing these predictions to the actual feedback, the starfish can thus exclude some of the possible models. When the robot has eliminated all but one model, a second phase starts and it uses this model to control its body and generate walking behavior (action as predictive control). Crucially, when the robot’s morphology changes (when an ex-

perimenter removes one of its limbs), it can switch back to the first phase, re-creating competing models and using action to eliminate most of them (action as hypothesis-testing).

Seth points out that the second phase, in which the robot walks around, suggests that the main purpose of predictive models is to control behavior effectively, regardless of how accurately it represents the world or the body (p. 15). In the first phase, by contrast, exploratory actions are conducted in order to learn something about the body, not to reach a goal involving its environment (ibid.). As noted above, such instances of action conform more to Helmholtzian than to cybernetic roots (action as hypothesis-testing).

What this shows is that action can fulfill different purposes—not just theoretically, but also in real applications. The robot starfish uses action in at least two ways. Drawing on the often-noted analogy between PP and scientific practice (cf. Gregory 1980), Seth explores further purposes of action. This leads to a distinction between three types of active inference (pp. 18f.). The first involves active sampling to confirm predictions derived from currently active models; the second is employed to seek evidence that would disconfirm currently held hypotheses; the third involves sampling in order to disambiguate between alternative hypotheses (p. 19).

Crucially, Seth does not elaborate much on the notion of falsification or disconfirmation. He relates disconfirmation to Bayesian surprise (which formalizes the extent to which new evidence leads to a revision of prior representations, cf. Baldi & Itti 2010). Accordingly, he characterizes seeking falsifying evidence in terms of maximizing Bayesian surprise. However, the paper quoted in this context, Itti & Baldi (2009) only investigates the hypothesis that surprising information attracts attention, not that subjects act to maximize surprise. Friston et al. (2012, p. 6) clarify the relation between FEP and maximization of Bayesian surprise:

The term Bayesian surprise can be a bit confusing because minimizing surprise per se (or maximizing model evidence) in-

4 “[...] we observe under constant own activity, and thereby achieve knowledge of the existence of a lawful relation between our innervations and the presence of different impressions of temporary presentations [Präsentabilien]. All of our willful movements through which we change the appearance of things should be considered an experiment, through which we test whether we have grasped correctly the lawful behavior of the appearance at hand, i.e. its supposed existence in determinate spatial structures.” (My translation)

volves keeping Bayesian surprise (complexity) as small as possible. This paradox can be resolved here by noting that agents expect Bayesian surprise to be maximized and then acting to minimize their surprise, given what they expect.

In the following section, I will clarify the notion of falsification, and discuss the ways in which it is used in PP. More specifically, I will illustrate various types of active inference by drawing a slightly broader analogy with theory of science. In particular, I will consider views put forward by Karl Popper and Thomas Kuhn, respectively. This will serve to help us get a handle on the general merits of confirmation and disconfirmation. Furthermore, both Popper's falsificationism and Kuhn's paradigm change can be related to aspects of predictive processing, which will hopefully lead to a better understanding of hypothesis-testing in PP. As a consequence, I invite Seth to provide a refined treatment of the relation between falsification and active inference.

3 Is the Bayesian brain Kuhnian or Popperian?⁵

The free-energy principle subsumes the Bayesian brain hypothesis⁶ (cf. Friston 2009, p. 294). According to this view, processing in the brain can usefully be described as Bayesian inference. This means that the brain implements a probabilistic model that is updated in light of sensory signals using Bayes' theorem. More specifically, the brain combines prior knowledge about hidden causes in the world with a measurement of likelihood describing how probable the observed (sensory) evidence is, given various possible hidden causes. The result is a distribution (posterior) that describes how probable various possible causes are, given the obtained evidence. The process of determining the pos-

terior is often called *model inversion*. In FEP, this type of inference is approximated using variational Bayes, which establishes the connection to predictive processing (cf. footnote 2 above). FEP can thus either be seen as a particular instance of the Bayesian brain hypothesis, or as a generalization.

As mentioned above, it is often pointed out that perceptions in PP are analogous to scientific hypotheses. The Bayesian brain is thus a hypothesis-testing brain (this analogy is also referred to in titles of papers by Jakob Hohwy, see Hohwy 2010, 2012). Thanks to active inference, the Bayesian brain performs an active kind of hypothesis testing. The three types of active inference distinguished by Seth assign a role to both confirmation and disconfirmation (falsification). This dual role of active inference is also emphasized by (Friston et al. 2012, p. 19):

The resulting active or embodied inference means that not only can we regard perception as hypotheses, but we could regard action as performing experiments that confirm or disconfirm those hypotheses.

Further exploration of the analogy to theory of science reveals a puzzle: as we will see, doubts can be raised regarding the idea that a theory gains merit when it is confirmed (or even regarding the very notion of theory confirmation). Does this mean that the Bayesian brain generates hypotheses in an unscientific way?

3.1 The Popperian Bayesian brain

3.1.1 Conceptual clarification: From naïve to sophisticated falsificationism

According to Popper, science advances mainly by seeking falsifying evidence. In fact, falsifiability is Popper's proposed solution to the demarcation problem, i.e., the problem of specifying the difference between science and pseudo-science. Scientific theories posit universal propositions (scientific laws) that can never be proven in a strict sense, because only finite observations can be made. The next observation could, in principle, always disconfirm a universal em-

⁵ It should be noted that Popper rejected interpretations of confirmation (or corroboration) in terms of probabilities (cf. Popper 2005[1934], ch. X), as well as Bayesian interpretations of probability theory (cf. Popper 2005[1934], ch. *XVII). Here, I only suggest that a useful analogy between Popper's theory of science and the Bayesian brain can be drawn.

⁶ Seth identifies PP and the Bayesian brain (cf. p. 1). I follow suit in this commentary.

pirical hypothesis. Hence, being verifiable cannot be a criterion for being scientific, because theories cannot be empirically verified (cf. Popper 2005[1934], pp. 16-17.). Conversely, it is possible to *falsify* a universal statement using a single empirical proposition:

Diese Überlegungen legen den Gedanken nahe, als Abgrenzungskriterium nicht die Verifizierbarkeit, sondern die *Falsifizierbarkeit* des Systems vorzuschlagen; [...] *Ein empirisch-wissenschaftliches System muß an der Erfahrung scheitern können.* (Popper 2005[1934], p. 17)⁷

Scientific theories thus cannot, according to Popper, be verified, but only falsified. However, when attempts to falsify a hypothesis have failed, we can say that the theory has been *corroborated*—which still means that the theory could be falsified in the future (cf. Popper 2005[1934], ch. X).

How can we apply these ideas to predictive processing? First, we have to find an analogy to scientific theories. I suggest that models can be treated analogously to theories, because in PP, predictions or hypotheses are derived from models and then compared to bottom-up signals. This also fits the way in which Seth describes the starfish example (namely in terms of model selection). What does it mean that a model is falsified in PP?

The question is not a trivial one, as there seems to be a crucial disanalogy between hypothesis-testing in Popper's sense and hypothesis-testing in the Bayesian brain. The reason why scientific theories are falsifiable is that they allow deriving hypotheses deductively. This means if a hypothesis is falsified, the theory is falsified as well. By contrast, hypotheses in the Bayesian brain are not deductively entailed by the models from which they are derived: the relation between model and hypothesis is *probabilistic* (the hypothesis is more or less probable, given the model). Hence, when a hypothesis or prediction elicits a large prediction error, this

does not falsify the model; rather, it calls for an update to the effect that the model becomes less likely. Furthermore, according to Popper, it does not make sense to say that such hypotheses are corroborated to a greater or lesser extent. For being corroborated means that attempts at falsification have failed. But if it is in principle impossible to falsify a hypothesis, then saying that it has been corroborated becomes empty—worse, such hypotheses are not even scientific hypotheses (cf. Popper 2005[1934], pp. 248-249.). This, then, constitutes the puzzle mentioned above: if hypotheses in PP are not falsifiable, does this mean the Bayesian brain is unscientific?

This conclusion—that no useful analogy to Popper's theory of science can be drawn—rests on a naïve understanding of falsification (as emphasized by Imre Lakatos, cf. Lakatos 1970).⁸ A closer look at the notion of falsification reveals that the analogy can be upheld. Furthermore, it helps us gain a better grasp of the notion of falsification in the context of PP.

First of all, we can note that in actual scientific practice, it is not the case that scientists attempt to falsify an isolated, single hypothesis—and then try to come up with a new theory when the hypothesis has been falsified. Rather, scientists often operate with different versions of a theory at the same time, or seek to find the best parameters for a model. The outcomes of an empirical study are then used to eliminate some of the different theories or parameter ranges. This has already been acknowledged by Popper (cf. 2005[1934], p. 63., fn. 10). As Thomas Nickles puts it:

According to Popper, at any time there may be several competing theories being proposed and subsequently refuted by failed empirical tests—rather like balloons being launched and then shot down, one by one. (2014)

The result of this falsification procedure is that some of the competing theories are eliminated. This can already be seen as a slight departure

⁷ “These considerations suggest proposing not verifiability, but falsifiability as a demarcation criterion; [...] An empirical-scientific system must be able to break down in the light of empirical evidence.” (My translation)

⁸ I am grateful to Thomas Metzinger for pointing me to Lakatos' work on falsificationism.

from what Imre Lakatos calls naïve falsificationism: for the elimination may be based on a comparison, not on an isolated falsification procedure. If some of the theories are in some sense better than the others (for instance, by making more empirical predictions, or by being less complex), then they can be preferred without having *independent* reasons to reject the eliminated theories. However, Popper's falsificationism is even more sophisticated.

Popper noted that there were no theory-neutral empirical propositions. Descriptions of empirical facts are not immediately given, they are based on observations and involve interpretations (cf. Popper 2005[1934], p. 84, fn. 32). This means it is always possible to add auxiliary hypotheses to a theory, and thereby make the theory compatible with seemingly falsifying evidence. As a consequence, when it comes to determining whether a theory is scientific or not, we cannot consider an isolated theory, but must assume a diachronic stance, in which we consider how a theory is modified in the light of new evidence. Such modifications (e.g., auxiliary hypotheses) increase the empirical content of the theory (cf. Lakatos 1970, p. 183). As Popper puts it:

Bezüglich der Hilfhypothesen setzen wir fest, nur solche als befriedigend zuzulassen, durch deren Einführung der 'Falsifizierungsgrad' des Systems [...] nicht herabgesetzt, sondern gesteigert wird; in diesem Fall bedeutet die Einführung der Hypothese eine Verbesserung: Das System verbietet mehr als vorher.⁹ (Popper 2005[1934], p. 58)

When confronted with evidence that contradicts predictions, we are thus never forced to reject the theory from which the prediction has been derived. We may always modify the theory. But this modification must not be *ad hoc*. Auxiliary hypotheses that only make the theory compatible with the evidence, without having any addi-

tional value (without allowing new predictions), are not scientific.

Lakatos (1970) emphasizes that this entails a refined notion of falsificationism. He calls this sophisticated falsificationism (or sophisticated *methodological* falsificationism). A theory can only be falsified in this "sophisticated" manner when it has been replaced by a theory that:

1. has more empirical content (makes new predictions), and
2. makes at least one prediction that is empirically corroborated (cf. Lakatos 1970, pp. 183-184.).

3.1.2 Sophisticated falsification in the Bayesian brain

Popper's sophisticated falsificationism¹⁰ can more easily be applied to predictive processing, because it does not require that we reject a model whenever its predictions yield large prediction errors. Instead, the model can be updated to achieve a better fit with the data. Furthermore, we find a counterpart for the insight that there are no theory-neutral observations: bottom-up signals are never treated as raw data, but as being (more or less) noisy. Hence, prediction errors are weighted by expected precisions. When the expected precision is extremely low, prediction errors will be attenuated. A low expected precision can thus be seen as analogous to an auxiliary hypothesis that makes the model compatible with otherwise contradicting evidence. What is more, it is not an *ad hoc* move, because the precision estimate itself is also constantly being updated in light of the evidence. Similarly, when a model generates a significant amount of prediction error, but is strongly supported by a higher-level model with high prior probability, a relatively high amount of prediction error may not lead to a major revision of the model.

¹⁰ Lakatos (1970) points out that Popper himself never made a sharp distinction between naïve and sophisticated falsificationism, but that he accepted the assumptions underlying sophisticated falsificationism, at least in parts of his work—whereas the person Karl Popper may have been more of a naïve than a sophisticated falsificationist.

⁹ "Regarding such auxiliary hypotheses we stipulate that we allow only those hypotheses for which the 'degree of falsifiability' of the system is not decreased, but increased; in this case the introduction of auxiliary hypotheses means an improvement: The system prohibits more than before." (My translation)

Model competition in PP can also be seen as an instance of sophisticated falsificationism. Competition need not be resolved by eliminating those models that yield the largest prediction errors (as in the starfish robot). Instead, it may be that some models make more specific *counterfactual* predictions. Indeed, this seems to be the main rationale behind active inference in FEP.

According to the formalization provided in [Friston et al. \(2012, p. 4\)](#), active inference involves minimizing the entropy of a counterfactual density. This density links future internal states and hidden controls to hidden states, which cause sensory states; hidden controls are hidden states that can be changed by action ([Friston et al. 2012, p. 3](#)). A density has low entropy, roughly, if it assigns high values to a relatively small subset of states, and low values to most other sets of states. Predictions based on a probability density with very low entropy can thus be made with a high level of confidence, because most other possibilities are more or less ruled out (due to the low values assigned to them by the density). Formally, this is reflected in the proposition that the negative entropy of the counterfactual density is a monotonic function of the precision of counterfactual beliefs ([Friston et al. 2012, p. 4](#)).

The entropy of the counterfactual density is minimized with respect to hidden controls. In effect, this is a selection process, in which a model (here: a counterfactual density) is selected that has minimal entropy. The other models are eliminated, because they have higher entropies. We can say they are falsified in the sense of sophisticated falsificationism (but not in the sense of naïve falsificationism).

Another way in which model competition can be resolved without naïve falsification can be illustrated by the famous “wet lawn” example (cf. [Pearl 1988](#)). Suppose you enter your garden and find that the lawn is wet. There are at least two models that can explain this: either your sprinkler has been on during the night or it has rained. Let us assume that both models are initially equally likely (i.e., they have the same prior probability). When you now observe that your neighbor’s garden is also wet, the rain

model is corroborated, because it makes the strong prediction that the neighbor’s lawn is wet (i.e., the conditional probability that the neighbor’s lawn is wet, given that it has rained, is high). The other model is not incompatible with this evidence, but it is not supported by it as much (because the conditional probability that the neighbor’s lawn is wet, given that your sprinkler has been on, is not as high). In other words, it has been explained away. As [Jakob Hohwy](#) puts it:

The Rain model accounts for all the evidence leaving no evidence behind for the Sprinkler model to explain. Even though the Sprinkler model did increase its probability in the light of the first observation, it seems intuitive right to say that its probability is now returned to near its prior value. The model has been explained away. ([2010, p. 137](#))

Explaining away is another example of sophisticated falsification. Even when two or more models are compatible with the evidence (and with each other), there can be reason to prefer one of them and reject the others.

The clarification in this section should have shown that there is more to falsification than just “disconfirming” a hypothesis, and that competition between models can be resolved in different ways, not only in the way exemplified by the starfish robot. Furthermore, different types of sophisticated falsificationism are part and parcel of predictive processing.

Does this mean that the Bayesian brain is Popperian? This conclusion would be premature. The above can at best show that there are many situations in which the Bayesian brain is a sophisticated falsificationist. But there may be situations in which not even sophisticated falsification is possible or necessary. In the following section, I will argue that predictive processing also has Kuhnian aspects.

3.2 The Kuhnian Bayesian brain

According to Kuhn, scientific research develops in different recurring phases. Most of the time,

scientists work within an established paradigm, in which implications of theories are explored and puzzles are solved (cf. [Kuhn 1962](#), ch. IV). In this phase, falsification or confirmation do not play a role:

Normal science does and must continually strive to bring theory and fact into closer agreement, and that activity can easily be seen as testing or as a search for confirmation or falsification. Instead, its object is to solve a puzzle for whose very existence the validity of the paradigm must be assumed. Failure to achieve a solution discredits only the scientist and not the theory. (cf. [Kuhn 1962](#), p. 80)

At some stage, however, there will be anomalies, i.e., empirical observations that cannot be explained within the current paradigm. When these anomalies accumulate, scientists will try to explore new concepts and methods. If, using new concepts and methods, previously unexplainable anomalies can be accounted for, a scientific revolution can result, through which a new paradigm is established. Kuhn shares the sophisticated falsificationist's insight that theories are never rejected in isolation:

[...] the act of judgment that leads scientists to reject a previously accepted theory is always based upon more than a comparison of that theory with the world. The decision to reject one paradigm is always simultaneously the decision to accept another, and the judgment leading to that decision involves the comparison of both paradigms with nature *and* with each other. (cf. [Kuhn 1962](#), p. 77)

This shows that Kuhn's theory is in some respects in line with sophisticated falsificationism—but he goes beyond it, in that he doubts that a paradigm that has been adopted instead of another is always better or closer to the truth. The reason for this is that he claims competing paradigms to be incommensurable (cf. also [Feyerabend 1962](#)), which means that they typically use radically different concepts and methods (cf.

[Oberheim & Hoyningen-Huene 2013](#), §1). A new paradigm that becomes dominant is thus not marked by being closer to the truth, but mainly by constituting a departure from the old paradigm (cf. [Kuhn 1962](#), pp. 170-171). This seems to entail that scientific progress need not be a process in which theories approximate the truth to an ever higher degree.

Can we find an analogon for such a transition from one paradigm to the other in predictive processing? Above, we saw that the sophisticated falsificationist assumes that scientific progress happens only when a theory makes new predictions, and thereby leads to the discovery of new states of affairs. This need not always be the case in the Bayesian brain. When a model is changed to minimize free-energy, this does not mean that the empirical content or predictive power has been increased. A particularly clear example of this can be found in perceptual phenomena like binocular rivalry.

In binocular rivalry (cf. [Blake & Logothetis 2002](#)), subjects are presented with two different images, one to the left eye, the other to the right eye, e.g., a face and a house. According to a predictive coding account put forward by [Jakob Hohwy](#), [Andreas Roepstorff](#) & [Karl Friston](#) (2008), the brain generates two main competing models of what the stimuli depict, one corresponding to the face, the other corresponding to the house. However, only one of these models is consciously experienced at any given time (although there can be intermittent phases in which subjects report seeing a mixture of both stimuli, i.e., parts of the house and parts of the face at the same time, but usually non-overlapping). This means that the brain will tend to settle into one of two classes of states (one corresponding to perceiving the house, the other to perceiving the face). Since each of the models can only account for part of the visual input, both cause a significant amount of prediction error (cf. [Hohwy et al. 2008](#), p. 691). Over time, the prior probability of the currently assumed model (house or face, respectively) will decrease, leading to a revision of the hypothesis, until the brain settles into a state corresponding to the other percept, at least temporarily (cf. [Hohwy et al. 2008](#), pp.

692–694).¹¹ The crucial difference between this and cases like the wet lawn example or model selection in the starfish robot is that neither of the two competing models is in any sense better than the other (in terms of empirical content, simplicity, predictive power, etc.).

We can recast binocular rivalry in terms of Kuhnian paradigm changes. If we liken each of the two models (house/face) to a paradigm, we can say that perceiving a single object in binocular rivalry corresponds to the phase of normal science, in which many phenomena (inputs) can be explained. After some time, however, there are anomalies (increasing prediction error), which leads to a scientific crisis in which new directions are explored (intermittent phase in which no unified percept is generated), until a new form of scientific practice becomes dominant (scientific revolution), and a new phase of normal science (temporarily stable perception) is reached. The transition from one percept to the other does not go along with increased veridicality: neither of the two percepts is closer to the truth than the other.¹² This may also support the cybernetic idea that internal models are used in the pursuit of homeostasis, not to approximate the truth (as also noted by [Seth this collection](#), p. 15).

There is another analogy between the Bayesian brain and Kuhn's theory of science. According to Kuhn, it is indeterminate whether an anomaly (an unexpected experimental result, for instance) is something that should be regarded as just another puzzle or as a reason to reject the whole paradigm:

¹¹ Two possible reasons why the probability of the currently assumed model decreases are offered by the authors: either there is a hyper-prior to the effect that the world changes (which is why a static hypothesis becomes less likely over time), or there are random effects that lead to multistability, such that neural dynamics switch from one basin of attraction to another (cf. [Hohwy et al. 2008](#), p. 692).

¹² In fact, it seems that the notion of incommensurability has been inspired by Gestalt switches (as in the perception of a Necker cube), which are very similar to phenomena like binocular rivalry. However, [Kuhn](#) explicitly pointed out that there is a crucial difference between a Gestalt switch and a paradigm change: “[...] the scientist does not preserve the gestalt subject's freedom to switch back and forth between ways of seeing. Nevertheless, the switch of gestalt, particularly because it is today so familiar, is a useful elementary prototype for what occurs in full-scale paradigm shift” (1962, p. 85). I am grateful to Sascha Fink for drawing my attention to this statement.

Excepting those that are exclusively instrumental, every problem that normal science sees as a puzzle can be seen, from another viewpoint, as a counterinstance and thus as a source of crisis. ([Kuhn 1962](#), p. 79)

If it is treated as a puzzle, it yields questions like: how can we account for this phenomenon within our established framework? If it is treated as a counterinstance, a more fundamental solution is needed. This is analogous to the fact that whether two models in predictive processing are compatible or not depends on (hyper)priors (cf. [FitzGerald et al. 2014](#), p. 2). When a hyper-prior has it that two models are incompatible, this can either lead to a competition, in which one of the models is eliminated, or it can lead to a revision of the hyper-prior. (Which of the two possibilities corresponds more to puzzle solving, and which to something more fundamental will depend on whether the lower-level models or the high-level prior initially have a higher probability.) This is illustrated by the RHI (rubber hand illusion).

In the RHI ([Botvinick & Cohen 1998](#)), the brain harbors two contradictory sensory models. According to the visual model, tactile stimulation occurs on the surface of the rubber hand. According to the proprioceptive model, the felt strokes occur at a different location (i.e., where the real hand is located). While there is, in and of itself, no contradiction between these models, it is likely that the brain has a prior that favors common-cause explanations of sensory signals. Relative to this prior, there is a tension between the models: they seem to indicate that the seen stroking and the felt touch occur at distinct locations, which is odd, because they occur synchronously (and the prior has it that synchronous effects have a common cause, which speaks against two distinct locations). As [Jakob Hohwy](#) puts it:

[...] we have a strong expectation that there is a common cause when inputs co-occur in time. This makes the binding hypothesis of the rubber hand scenario a better explainer, and its higher likelihood

promotes it to determine perceptual inference and thereby resolve the ambiguity. (2013, p. 105)

Notice that the common-cause hypothesis (that the touch is felt where it is seen) only becomes the dominating hypothesis because the design of the study prevents subjects from confirming the distinct-causes hypothesis (e.g., by looking at their real hands). Because of the common-cause hypothesis, there is an ambiguity in the percepts. This ambiguity can be resolved in at least two ways: either by adjusting the lower-level (perceptual) models (to the effect that the felt touch occurs at the same location as the seen stroking); or by active inference (which in this case would lead to a rejection of the higher-level model corresponding to the common-cause hypothesis). The first way corresponds to puzzle solving, the second more closely to a paradigm change. Note that the analogy will be the stronger the more remote the hyper-prior is from the perceptual models.

I hope to have shown that the Bayesian brain has aspects that make it Popperian, as well as aspects that make it Kuhnian. At the very least, it should have become clear that falsification is a more complex concept than depicted in Seth's target paper (which seems to tend towards a more naïve form of falsificationism).

4 Perceptual presence

We have seen how fruitful analogies between PP and theory of science can be. As mentioned above, an early formulation of the analogy between perception and hypothesis-testing can be found in Richard Gregory's seminal paper "Perceptions as Hypotheses". There, we also find the suggestion that percepts *explain* sensory signals (cf. Gregory 1980, p. 13).¹³

How far can we take the analogy between explanation in perception and explanation in science? If we know what a good explanation is in science, does this give us a clue to the conditions under which percepts are experi-

enced as real? Interestingly, there are accounts of scientific explanation that assign an essential role to counterfactual knowledge (cf. Waskan 2008). If someone purports to know why a certain event happened or why a phenomenon was observed, we expect her to also be able to tell us what *would* have happened if some of the initial conditions had been different. Similarly, when the Bayesian brain explains sensory signals by inferring their hidden causes, we would expect the brain's generative model to also have the resources to infer in what ways sensory signals would be different, had there been a change to their hidden causes.

This highlights the relevance of counterfactual models. Seth points out that counterfactuals play a crucial role in active inference. The consideration above may be another way to show the relevance of counterfactual models. Furthermore, it also highlights the usefulness of counterfactual *richness*. The richer a counterfactual model of hidden causes, the better the brain's explanation of sensory signals (all other things being equal). In general, we may also be inclined to say that the richer the counterfactual model, the higher the confidence that it helps track the *real* explanation of sensory signals. But does this mean it goes along with experienced *realness* (or *perceptual presence*)?

This is, basically, what Seth proposes in his PP account of perceptual presence (cf. Seth 2014). But what is perceptual presence in the first place? On the one hand, Seth characterizes the notion by contrasting examples. For instance, objects like a tomato possess perceptual presence, whereas afterimages do not. On the other hand, Seth provides the following characterization:

In normal circumstances perceptual content is characterized by subjective veridicality; that is, the objects of perception are experienced as real, as belonging to the world. When we perceive the tomato we perceive it as an externally existing object with a back and sides, not simply as a specific view [...]. (2014, p. 98)

¹³ It should be noted that Gregory ascribes "far less explanatory power" (1980, p. 196) to perceptions than to scientific hypotheses.

The tomato is not perceived as a flat, red disc. Although you do not see the back and sides of the tomato in the same way that you see the front, there is still a sense in which both are *perceptually present* (cf. Noë 2006, p. 414). I shall now point to two ambiguities in Seth's description of the explanandum. This calls for a conceptual clarification, regarding which I shall make a tentative suggestion. After that, I shall argue that there may be possible counterexamples to Seth's hypothesis that perceptual presence correlates with the counterfactual richness of generative models.

4.1 Ambiguities in Seth's description of the explanandum

The tomato is not only experienced as perceptually present, it is also perceived as an *object* in the external world. In a commentary on Seth, Tom Froese (2014, p. 126) has therefore suggested that Seth conflates perceptual presence with experienced *objecthood*. This proposal has some plausibility, because the tomato is perceived as a real object, whereas afterimages are not experienced as objects (they are more like unstable colored shades). After all, even Seth admits, in his target paper, that it may be important to distinguish presence from objecthood (p. 18). This is one way in which Seth's definition of the explanatory target is ambiguous: is it about experienced *presence* or experienced *objecthood* (cf. also Seth 2014, pp. 105f.)? (This question becomes more pressing still when we consider the etymology of "realness" or "reality": the Latin origin of the word is *res* (thing), which makes it a little confusing that Seth seems to identify perceptual presence with the sense of subjective reality, cf. Seth this collection, p. 2.)

Another ambiguity is related to the notion of a counterfactual model. In his target paper Seth defines a counterfactual model as a model encoding "how sensory inputs (and their expected precisions) would change on the basis of a repertoire of possible actions" (Seth this collection p. 17). On the one hand, one may ask if counterfactual models in the brain necessarily encode SMCs (sensorimotor contingencies). For

the perception of a ripe tomato on a bush, it might be equally relevant to encode how sensory signals pertaining to the tomato would change if the wind were to blow the bush or if the tomato were to fall down. On the other hand, it is unclear how *explicit* a counterfactual representation has to be. Jakob Hohwy (2014) suggests that a rich causal structure could be modeled by extracting higher-order invariants (features that do not change if the tomato is dangling in the wind or has fallen down, for instance). Higher-order invariants are relatively perspective-independent.¹⁴ The degree of perceptual presence would then correspond to the "depth of the inverted model"¹⁵ (Hohwy 2014, p. 128). In his target paper, Seth notes that the depth of the model may indeed play a role (see footnote 13).

Two ambiguities are thus to be found in Seth's account. One concerns the characterization of the target phenomenon (experienced *realness* versus experienced *objecthood*). The other lies in the description of the represented causal structure: *counterfactual richness* versus *perspective-independence* of hidden causes. Counterfactual richness and causal "depth" are not completely independent. Below, I will give some examples that may be useful to explore the relationship between these two features. Furthermore, I will suggest that it could be helpful to consider another feature with respect to which the represented causal structure of objects may vary. This feature is the degree of

¹⁴ As I am using the term here, the depth of a model can be measured by its location in the predictive processing hierarchy (that is, whether it is high or low in the hierarchy). Estimates at higher levels track features that change more slowly (i.e., features that remain invariant when things change, for instance, when the subject changes her *perspective* on a perceptual object like a tomato by walking around the tomato or by turning it—hence the term "perspective-(in)dependence"). A model of a perceived object is deep when it represents features that change relatively slowly. Alternatively, one could stipulate that a model is deep when it represents features that change slowly *and* features that change more quickly. In fact, this may come closer to what Hohwy has in mind, but it blurs the distinction between perspective-dependence and causal integration. Hohwy writes: "[c]oncurrents are causes that do not interact on their own with other causes (presumably a fence won't occlude a concurrent)" (2014, p. 128). But encapsulated causes can be represented both at lower parts of the hierarchy (possible example: afterimages) and at higher parts of the hierarchy (possible example: certain conscious thoughts). This suggests that at least causal encapsulation can be dissociated from perspective-dependence and -independence.

¹⁵ The inverted model is the posterior distribution, the computation of which is based on the likelihood and the prior (see above).

causal encapsulation. For representations not only differ with respect to their counterfactual richness or their degree of perspective-dependence, but also with respect to the extent to which the represented causal structure is encapsulated or integrated. (In what follows, I will use the notion of a counterfactual model mainly in the sense in which Seth uses it: counterfactual models in this sense involve representations of possible bodily actions by the subject of experience.)

A phenomenal representation of a tomato on a plate is not only counterfactually rich and relatively perspective-dependent, the represented causal structure is also causally *integrated*.¹⁶ It is, for instance, represented as being causally related to the plate, because it is experienced as lying *on* the plate (that is, it is not hovering above it). Furthermore, it is in possible causal contact with virtually all other objects in its vicinity (e.g., the subject's hands).

Contrast this with the experience of what is happening in a classical video game—say, a racing game. The player influences how the images on the two-dimensional screen change, because she has control over the vehicle. Hence, we can assume that representations of gaming sequences are (usually) counterfactually rich. At the same time, they are also perspective dependent (although they mainly depend on the *virtual* perspective from which objects are represented in the game). However, virtual objects in the game are experienced as causally encapsulated: although objects can interact with each other in the virtual world, they do not interact with most other parts of the player's environment. For instance, they will never break out of the screen and fly around in the room in which the player is sitting. Furthermore, they can only be influenced vicariously through a controller or keyboard. Thus there is not causal encapsulation in *every* respect (the virtual world is not experienced as completely disambiguated from the rest of the experienced world), but in *some* respects the encapsulation is rather strong (the

virtual world is spatially bounded, e.g., with the screen as the limit). Note that many modern video games are less causally encapsulated, for instance when they are played on a touchscreen (or on devices with a three-dimensional screen, or in an immersive virtual reality).¹⁷

As mentioned above, causal integration and counterfactual richness are not completely independent. High counterfactual richness implies a certain degree of causal integration (at least in some respects), for it means that at least the subject can interact with the experienced object in some way—regardless of how separate the represented causal structure is from the rest of the subject's surroundings.

Similarly, highly perspective-invariant representations typically also involve the representation of an encapsulated causal structure. Abstract conscious thoughts, for instance, cannot be touched with the hand or other concrete objects. However, the implied encapsulation only holds in some respects. Sometimes thoughts can evoke strong emotions or a sequence of mental imagery. In certain obsessive-compulsive disorders, for instance, subjects will first have a thought (“My hands are dirty”), presumably followed by a feeling of disgust and the urge to wash the hands, which then leads to motor behavior (washing the hands); this, in turn, may be followed by the thought that the hands are still dirty. The content of the conscious thought is relatively perspective-invariant, and yet it involves, presumably, representations of causal structure that link it to concrete objects in the world.

As long as we interpret counterfactuals only as representations of sensorimotor contingencies, it may also seem that perspective-invariant¹⁸ representations are counterfactually poor. However, if we include representations of possible *mental* actions and their effects, we can also conceive of counterfactually-rich perspective-invariant representations. A possible example is a philosophical argument or a theory, which someone can contemplate in their mind, being aware that there are several possible ways

¹⁶ Another possible term for this would be *causally open*, in the sense that it is represented as being in potential causal exchange with other objects in its surrounding. By integration, I thus do not mean integration *within* (or internal integration), but integration *with* other objects.

¹⁷ Thanks to Jennifer Windt for suggesting immersive video games as a further example.

¹⁸ Perspective-invariant representations are maximally perspective-independent.

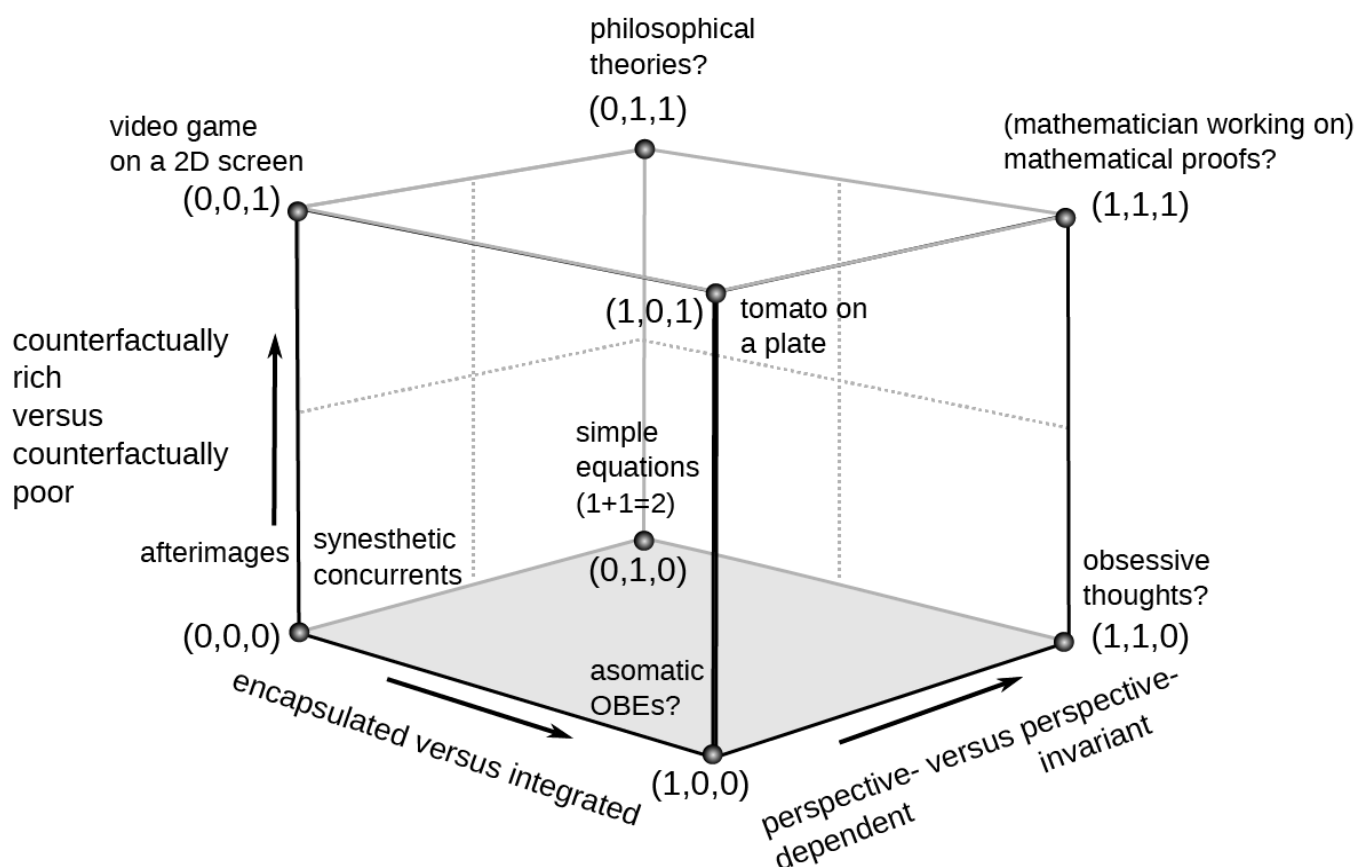


Figure 1: The figure illustrates how classes of experiences can be located in a cube, according to the extent to which they display counterfactual richness, perspective-independence, and causal integration (see main text for explanations). The cube (without the labels) is adapted from cube figures in [Godfrey-Smith \(2009\)](#); talks by Daniel Dennett brought this style of illustration to my attention.

in which the argument could be probed and attacked, or several important cases to which the theory could be applied.

Bearing in mind that the degree of causal encapsulation is not completely independent from the other two dimensions (counterfactual richness and perspective-invariance), we can depict different types of conscious experiences in a cube, where the three axes stand for the three dimensions described (see [Figure 1](#)). The most interesting locations in this cube are, of course, its eight corners, because they depict classes of experiences for which each of the three features is either completely absent or maximally pronounced. Finding examples of these “extremal experiences” is no easy task.¹⁹ Even neural representations of synesthetic concurrents, Seth’s prime example of coun-

terfactually poor models, may, at first sight, seem to be located somewhere in the middle of the perspective-dependence axis.

Grapheme-color concurrents, for instance, are not simply triggered by graphic representations of glyphs, but by representations of abstract objects, i.e., graphemes, associated with certain glyphs (cf. [Mroczko et al. 2009](#)). Hence, it may seem that the hidden cause of the concurrent is not simply an object in the world, but also involves an abstract object, i.e., a grapheme, the representation of which is perspective-invariant. This would suggest that synesthetic concurrents cannot conclusively be placed in one of the cube’s corners, because their represented hidden causes involve very high-level invariants.

On the other hand, one could object that the concurrent itself is represented in a rather perspective-dependent way. It may be part of a

¹⁹ In fact, it may be that the corners only constitute hypothetical endpoints. Thanks to Jennifer Windt for pointing this out.

causal network involving hidden causes that are represented in perspective-invariant ways, but the synesthetic percept itself is not a representation of an abstract hidden cause.²⁰ Hence, on second thought, it seems that concurrents, as in grapheme-color synesthesia, are in fact located close to the origin of our coordinate system: the representations involved are relatively perspective dependent, and they are counterfactually poor. At the same time, they are causally encapsulated, because they do not interact with physical objects (they cannot be touched, etc.).

4.2 Does counterfactual richness correlate with perceptual presence (or objecthood)?

What does this tell us about experienced “presence” or “objecthood”? Are all examples of counterfactually rich representations in the cube perceptually present, or are they associated with a high degree of objecthood? If so, this would support Seth’s hypothesis that counterfactual richness correlates with perceptual presence (or objecthood). I believe that counterfactual richness can be dissociated both from perceptual presence and from objecthood. Olfactory experiences are, as argued by Michael Madary (2014), both counterfactually poor and perceptually present. This suggests that counterfactual richness does not correlate with perceptual presence. Similarly, experiences of classical video game sequences are counterfactually rich, but involve a low degree of perceptual presence; objects in the game are only experienced as virtual objects, not as real objects. Counterfactual richness and perceptual presence may therefore be doubly dissociable.

Trying to evaluate whether counterfactual richness correlates with phenomenal objecthood would presuppose that we know what phenomenal objecthood means. As I only have an intuitive grasp of what it means, I can only give a preliminary statement. To me, it seems that virtual objects in two-dimensional video games do not possess a high degree of phenomenal objecthood. But then again, even if a virtual tomato

could be manipulated in various ways with a controller, the corresponding representation would probably not be as counterfactually rich as a representation corresponding to the experience of a real tomato. Hence, it is difficult to arrive at a definitive verdict.

A more promising path may involve the experience of objects in asomatic OBEs (out-of-body experiences) or asomatic dream experiences (Windt 2010; Metzinger 2013). Counterfactuals, as conceived of by Seth, always involve action on the part of a subject. Most, if not all, (non-mental) actions involve the body, so representing counterfactuals involves representing (parts of) the body. In asomatic OBEs and asomatic dream experiences, subjects do not identify with a body, but with an unextended point in space. I speculate that in such cases, representations of objects are less counterfactually rich.²¹ This, however, does not necessarily mean that they are experienced as less present or as possessing less objecthood. There are still a lot of causal regularities involving external objects that may be tracked by models in the brain, even in the absence of an ordinary body representation. External objects can interact with each other, and counterfactual representations of possible causal processes may contribute to the experience of objecthood or perceptual presence. In particular, this is to be expected if none of the external objects are represented as causally encapsulated. If this bears out, it provides another reason to believe that counterfactual richness of generative models does not correlate with experienced objecthood. Let us now consider possible examples of other extremal experiences (in the corners of the cube) to investigate whether it is plausible to hypothesize that represented causal depth or causal encapsulation correlates with perceptual presence or objecthood.

The more perspective-invariant a representation, the more abstract it is. This also means that perspective-invariant representations typically involve an encapsulated causal structure. Thinking about a simple equation like

²⁰ This may point to an aspect regarding which Hohwy’s characterization of causal depth is ambiguous.

²¹ In fact, asomatic OBEs may be a better example than asomatic dream experiences, since such dreams typically lack concrete objects (cf. LaBerge & DeGracia 2000). I am grateful to Jennifer Windt for pointing this out.

“ $1+1=2$ ” may be an example of this. There is no way in which the target of this representation can causally interact with the window behind my desk or the red bottle in front of the window. Furthermore, most (or all) bodily movements will not influence the way I experience the thought that one plus one equals two. Hence, it is arguably also a counterfactually poor representation.

When we move up, in the direction of counterfactually rich phenomenal representations, we arrive at representations that are counterfactually rich, perspective-invariant, and still causally encapsulated. Above, I mentioned conscious thoughts about philosophical arguments or theories as possible examples. Such thoughts may involve mental imagery and inner speech, and perhaps even complex phenomenal simulations involving counterfactual situations. It is not obvious whether it makes sense to say that such thoughts involve counterfactual representations linking possible mental actions to their effects. This is even harder without presupposing a developed theory of mental action (for recent proposals, cf. Proust 2013; Wu 2013).

Mental actions are goal-directed. Performing a mental action may therefore, at least in some cases, be followed by a representation of a situation in which the goal is realized (one possible example might be: remembering a name; represented situation: telling someone the name). In the case of a theory, a mental action could be considering whether a certain claim is true or not (or whether it is plausible). This may trigger thoughts like: “Assuming this is the case, what implications would this have? Are these implications plausible, or likely to be true? Are there possible counterexamples?” It might also involve trying to formulate something more clearly.

Furthermore, thinking about a theory or problem may involve conscious counterfactual thoughts of the form “If I gave up this assumption, there would not be a contradiction among the remaining hypotheses anymore”, or “If the theory could account for this special case, it would be strengthened”. One difference to conscious perception of concrete objects is, presum-

ably, that such counterfactuals are *phenomenally* represented, whereas representations of SMCs are usually unconscious (and may impact on consciousness only indirectly).

Similar things apply to conscious thoughts about non-trivial mathematical expressions. For instance, if a mathematician sees the expression $(1 + x/n)^n$ she will probably think “If n tends to infinity, this expression will converge to e^x ”. Now, suppose the mathematician is investigating the asymptotic behavior of some complicated expression (e.g., she wants to find out what happens to a certain expression when n tends to infinity). While manipulating the terms on paper, she suddenly realizes that one factor contained in the expression is $(1 + x/n)^n$. As she is using pen and paper while thinking this, her brain will not only activate an abstract (but conscious) counterfactual thought, but probably also a representation of SMCs. These SMCs will involve taking the limit of the expression with which she started (i.e., $\lim_{n \rightarrow \infty}$), and this is now not only a mental action, but also a possible bodily action. She could write this down, and know that (if the limit exists) part of it would be e^x . Her mathematical investigation therefore involves:

- phenomenal representations regarding counterfactual mental actions;
- representations of SMCs (*embodied* versions of the above mentioned counterfactuals);
- a close coupling between writing, perceiving, and thinking.

The third point is especially important, because it suggests that for a mathematician working with pen and paper (or chalk and blackboard) the objects of her conscious thoughts are not causally encapsulated anymore. The causal structure represented while thinking about abstract concepts is intertwined with the causal structure represented while looking at written mathematical expressions. These causal relations are still relatively limited, but if the mathematician is completely absorbed in her work, the paper (or blackboard) may be all she is attending to in her environment at the moment, perhaps to the extent that she does not experi-

ence abstract relations represented by her notes as causally encapsulated anymore. It is conceivable that this aspect can be enhanced in virtual environments in which mathematical objects are not represented by writing on paper or blackboard, but by three-dimensional virtual objects that can be manipulated by touch or manual movements, for instance.²² Contrary to what one might at first think, there may thus be cases in which high-degrees of perspective-invariance go along with both counterfactual richness and high degrees of causal integration.

Another class of abstract thoughts that may be experienced as causally integrated could be obsessive thoughts, like the thought that one's hands are contaminated with germs. Such thoughts may be triggered by specific events (like touching a door knob) and may go along with a fear of getting sick (because of the contamination). Subjects may also try to avoid touching objects that they fear might be contaminated. The reason for this is that the hidden cause represented by the obsessive thought, i.e., potential germ contamination, is not causally encapsulated. It is causally connected to concrete objects in the subjects' environment: things that are perceived as contaminated can cause a contamination of the hands; on the other hand, contaminated hands can infect other objects with germs. Furthermore, the inferred hidden cause (germ contamination) is relatively perspective-invariant. Subjects arguably do not imagine bacteria crawling on their hands, although the obsessive thought may go along with an altered perception of the hands. Finally, the model involved is probably counterfactually poor, as most actions do not change the alleged contamination (with the possible exception of washing the hands or touching allegedly contaminated objects; but here, the counterfactual effect is probably just an increase or decrease in the acuteness of the felt contamination). Therefore, I list obsessive thoughts as candidate examples of counterfactually poor, perspective-invariant representations the contents of which are represented as causally integrated.

4.3 Do perspective-invariance or represented causal integration correlate with perceptual presence (or objecthood)?

The examples given are certainly not uncontroversial and perhaps not all of them can be sustained in the light of further research. But hopefully the cube can still fulfill heuristic purposes, and can illustrate the need to clarify the relations between counterfactual richness, perspective-dependence, and causal integration. But assuming that the examples given are located in roughly the right places within the cube, what does this tell us about perceptual presence or experienced objecthood? Above, I dismissed Seth's hypothesis that counterfactual richness correlates with either presence or objecthood. Let us now briefly consider perspective-invariance and causal integration. If conscious thoughts involve causally-deep models (that represent perspective-invariant features), then it seems that the depth of the represented causal structure does not correlate with presence or objecthood. The thought that one plus one equals two does not possess a high degree of objecthood or perceptual presence. Hence, it seems that Hohwy's hypothesis that the depth of the generative model (the degree of perspective-independence) correlates with objecthood or presence should be dismissed as well. But the remaining candidate, causal integration, does not unequivocally correlate with either presence of objecthood (*if* the examples I gave make sense). The represented causal structure in obsessive thoughts need not be encapsulated, and still they are probably not accompanied by experienced objecthood or perceptual presence. Perhaps this shows that one ought first to clarify whether it even makes sense to talk about the phenomenology of objecthood or presence with respect to conscious thoughts.

4.4 How does perception change when new sensorimotor contingencies are learnt?

Another relevant question is whether increasing the degree of counterfactual richness, causal integ-

²² This could be a case in which there is a particularly strong demand for the general ability of PP to combine "fast and frugal solutions" with "more structured, knowledge-intensive strategies" (Clark [this collection](#)).

ration, or causal depth of a model just modifies (or enriches) the inferred hidden causes, or whether it leads to the perception of a new, possibly more abstract object. This relates to the question raised in the target paper, namely whether a person who is highly familiar with an object perceives it as more real (because she has mastery of more SMCs) than other persons (Seth this collection, p. 18). Interestingly, research on learning new SMCs tentatively suggests that it leads to the perception of new (more abstract) objects.

Under the lead of Peter König, cognitive scientists from Osnabrück have, in recent years, developed a compass belt that indicates to the person wearing it (while moving) changes in directions (cf. Kaspar et al. 2014). The aim of this project (called *feelspace*) is to study how perception in new sensory modalities can be enabled by sensory augmentation.²³ The belt (see Figure 2) contains several vibrators, which always signal the direction of magnetic north. Subjects who wear the belt for a couple of weeks learn new SMCs, e.g., related to how the vibrating signals change when they turn around. A straightforward application of Seth's PPSMCT suggests that the increased counterfactual richness simply goes along with an increased perceptual presence (for the belt, or the vibrations, or the hip / waist, etc). But the authors of the study cited report that perception changes in different ways:

Initially the signal was predominantly perceived as tactile evolving to being perceived as location and direction information. Over time, the perception of tactile stimulation receded more and more into the background. Instead the subjects' reports focused more on changes in spatial perception. Furthermore, two months after the end of belt wearing the effects subjects reported – at least in the FRS questionnaire – diminished. (Kaspar et al. 2014, p. 59)

What changes is not just that SMCs for tactile stimulation on the skin where the belt is worn are learnt, but that these are connected to

more abstract information (regarding location and direction). This also makes sense in comparison with other sensory modalities. Knowledge of auditory SMCs, for instance, does not increase the perception of the inner ear. When the brain learns the relevant SMCs, it thereby learns about the hidden causes of signals in the inner ear. In fact, this may be another reason to believe that counterfactual richness goes along with phenomenal objecthood.

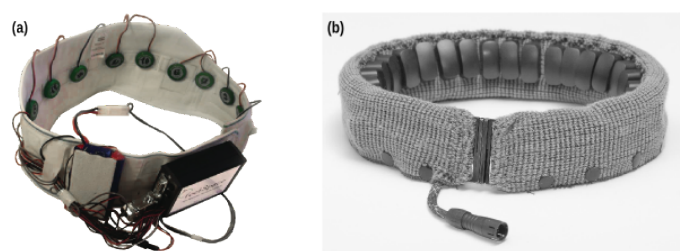


Figure 2: The figure shows two versions of the feelspace belt. (a) The original version used in Nagel et al. (2005). (b) The current version used in Karcher et al. (2012) and Kaspar et al. (2014). Images used with kind permission of Peter König.

This also suggests that when someone is more familiar with an object, the object itself need not become more real, but its connections to other objects might. The causal network in which it is embedded becomes more real. Perhaps the subject also experiences more abstract objects (corresponding to higher-level invariants).

All in all, I hope the examples given illustrate the need to provide a conceptually clearer account of counterfactual richness, causal depth, and causal integration. For at the moment it seems that they are too entangled to allow us to assess their potential relevance for experienced objecthood or presence in a rigorous way. Furthermore, it will be crucial to investigate how phenomenal properties are affected when there are *changes* in these three features (e.g., when counterfactual richness or causal integration is increased or decreased in a controlled way in a study).

5 Conclusion

I have tried to show that useful analogies between PP accounts and classical ideas in the-

²³ For more information on the project, see: <http://feelspace.cogsci.uni-osnabrueck.de/>

ory of science run deeper than portrayed in Seth's target paper. Based on such analogies, I have argued that a proper treatment of active inference needs to be more sophisticated than Seth's threefold distinction. In particular, Seth blurs a whole range of ways in which models can be falsified.

Furthermore, I have suggested that Seth's predictive processing account of perceptual presence may profit from taking not just the counterfactual richness of generative models, but also their degree of perspective-dependence and their causal encapsulation into account (as mentioned above, this suggestion is inspired by Jakob Hohwy's work). I have proposed a way in which examples of possible combinations of these features can be explored, which may serve as a useful tool for future research.

Thomas Kuhn (1962, p. 88) writes that "normal science usually holds creative philosophy at arm's length, and probably for good reasons". I thus hope that research on predictive processing and consciousness has not yet reached the phase of normal science, so that this commentary can still make a humble contribution.

Acknowledgments

I am grateful to two anonymous reviewers, and to Jennifer Windt and Thomas Metzinger especially for providing a vast number of comments and remarks, which helped tremendously in revising the first draft of this paper. This comment was written with support by a scholarship from the Barbara Wengeler foundation.

References

- Ashby, W. R. (1947). Principles of the self-organizing dynamic system. *The Journal of General Psychology*, 37 (2), 125-128. [10.1080/00221309.1947.9918144](https://doi.org/10.1080/00221309.1947.9918144)
- Baldi, P., & Itti, L. (2010). Of bits and wows: A Bayesian theory of surprise with applications to attention. *Neural Networks*, 23 (5), 649-666. [10.1016/j.neunet.2009.12.007](https://doi.org/10.1016/j.neunet.2009.12.007)
- Blake, R. & Logothetis, N. K. (2002). Visual competition. *Nature Reviews Neuroscience*, 3 (1), 13-21.
- Bongard, J., Zykov, V., & Lipson, H. (2006). Resilient machines through continuous self-modeling. *Science*, 314 (5802), 1118-1121. [10.1126/science.1133687](https://doi.org/10.1126/science.1133687)
- Botvinick, M. & Cohen, J. (1998). Rubber hands 'feel' touch that eyes see. *Nature*, 391 (6669), 756-756. [10.1038/35784](https://doi.org/10.1038/35784)
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36 (3), 181-204. [10.1017/S0140525X12000477](https://doi.org/10.1017/S0140525X12000477)
- (2015). Embodied Prediction. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1-21). Frankfurt a. M., GER: MIND Group.
- Feyerabend, P. (1962). Explanation, reduction and empiricism. In H. Feigl & G. Maxwell (Eds.) *Scientific explanation, space, and time* (pp. 28-97). Minneapolis, MN: University of Minnesota Press.
- FitzGerald, T. H., Dolan, R. J., & Friston, K. J. (2014). Model averaging, optimal inference and habit formation. *Frontiers in Human Neuroscience*, 8 (457), 1-11. [10.3389/fnhum.2014.00457](https://doi.org/10.3389/fnhum.2014.00457)
- Friston, K. J. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, 13 (7), 293-301. [10.1016/j.tics.2009.04.005](https://doi.org/10.1016/j.tics.2009.04.005)
- (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11 (2), 127-138. [10.1038/nrn2787](https://doi.org/10.1038/nrn2787)
- Friston, K. J., Adams, R., Perrinet, L., & Breakspear, M. (2012). Perceptions as hypotheses: saccades as experiments. *Frontiers in Psychology*, 3 (151), 1-20. [10.3389/fpsyg.2012.00151](https://doi.org/10.3389/fpsyg.2012.00151)
- Froese, T. (2014). Steps toward an enactive account of synesthesia. *Cognitive Neuroscience*, 5 (2), 126-127. [10.1080/17588928.2014.905521](https://doi.org/10.1080/17588928.2014.905521)
- Godfrey-Smith, P. (2009). *Darwinian populations and natural selection*. Oxford, UK: Oxford University Press.
- Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 290 (1038), 181-197. [10.1098/rstb.1980.0090](https://doi.org/10.1098/rstb.1980.0090)

- Hohwy, J. (2010). The hypothesis testing brain: some philosophical applications. In W. Christensen, E. Schier & J. Sutton (Eds.) *Proceedings of the 9th Conference of the Australasian Society for Cognitive Science* (pp. 135-144). Macquarie Centre for Cognitive Science. [10.5096/ASCS200922](https://doi.org/10.5096/ASCS200922)
- (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology*, 3 (96), 1-14. [10.3389/fpsyg.2012.00096](https://doi.org/10.3389/fpsyg.2012.00096)
- (2013). *The predictive mind*. Oxford, UK: Oxford University Press.
- (2014). Elusive phenomenology, counterfactual awareness, and presence without mastery. *Cognitive Neuroscience*, 5 (2), 127-128. [10.1080/17588928.2014.906399](https://doi.org/10.1080/17588928.2014.906399)
- Hohwy, J., Roepstorff, A. & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, 108 (3), 687-701. <http://dx.doi.org/10.1016/j.cognition.2008.05.010>
- Itti, L. & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, 49 (10), 1295 – 1306. <http://dx.doi.org/10.1016/j.visres.2008.09.007>
- Kärcher, S. M, Fenzlaff, S., Hartmann, D., Nagel, S. K., & König, P. (2012). Sensory augmentation for the blind. *Frontiers in Human Neuroscience*, 6 (37), 1-15. Frontiers Media SA. [10.3389/fnhum.2012.00037](https://doi.org/10.3389/fnhum.2012.00037)
- Kaspar, K., König, S., Schwandt, J., & König, P. (2014). The experience of new sensorimotor contingencies by sensory augmentation. *Consciousness and Cognition*, 28, 47-63. [10.1016/j.concog.2014.06.006](https://doi.org/10.1016/j.concog.2014.06.006)
- Kuhn, T. S. (1974). *The structure of scientific revolutions*. Chicago, IL: The University of Chicago Press.
- LaBerge, S. & DeGracia, D. J. (2000). Varieties of lucid dreaming experience. In R. G. Kunzendorf & B. Wallace (Eds.) *Individual differences in conscious experience* (pp. 269-307). Amsterdam, NL: John Benjamins.
- Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & Musgrave, A. (Eds.) *Criticism and the growth of knowledge* (pp. 91-196). Cambridge, UK: Cambridge University Press.
- Madary, M. (2014). Perceptual presence without counterfactual richness. *Cognitive Neuroscience*, 5 (2), 131-132. [10.1080/17588928.2014.907257](https://doi.org/10.1080/17588928.2014.907257)
- Metzinger, T. K. (2013). Why are dreams interesting for philosophers? The example of minimal phenomenal selfhood, plus an agenda for future research. *Frontiers in Psychology*, 4 (746). [10.3389/fpsyg.2013.00746](https://doi.org/10.3389/fpsyg.2013.00746)
- Mroczko, A., Metzinger, T., Singer, W., & Nikolić, D. (2009). Immediate transfer of synesthesia to a novel inducer. *Journal of Vision*, 9 (12), 1-8. [10.1167/9.12.25](https://doi.org/10.1167/9.12.25)
- Nagel, S. K., Carl, C., Kringe, T., Martin, R., & König, P. (2005). Beyond sensory substitution--learning the sixth sense. *Journal of Neural Engineering*, 2 (4), 13-26. [10.1088/1741-2560/2/4/R02](https://doi.org/10.1088/1741-2560/2/4/R02)
- Nickles, T. (2014). Scientific revolutions. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. <http://plato.stanford.edu/entries/scientific-revolutions/>
- Noë, A. (2006). Experience without the head. In T. S. Gendler & J. Hawthorne (Eds.) *Perceptual experience* (pp. 411-434). Oxford, UK: Oxford University Press.
- Oberheim, E. & Hoyningen-Huene, P. (2013). The incommensurability of scientific theories. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. <http://plato.stanford.edu/entries/incommensurability/>
- Pearl, J. (1988). Embracing causality in default reasoning. *Artificial Intelligence*, 35 (2), 259-271. [10.1016/0004-3702\(88\)90015-X](https://doi.org/10.1016/0004-3702(88)90015-X)
- Popper, K. R. (2005[1934]). *Logik der Forschung*. Tübingen, GER: Mohr Siebeck.
- Proust, J. (2013). Mental acts as natural kinds. In A. Clark, J. Kiverstein & T. Vierkant (Eds.) *Decomposing the will* (pp. 262-280). Oxford, UK: Oxford University Press.
- Seth, A. K. (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive Neuroscience*, 5 (2), 97-118. [10.1080/17588928.2013.877880](https://doi.org/10.1080/17588928.2013.877880)
- (2015). The Cybernetic Bayesian Brain. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1-25). Frankfurt a. M., GER: MIND Group.
- Von Helmholtz, H. (1959). *Die Tatsachen in der Wahrnehmung. Zählen und Messen*. Darmstadt, GER: Wissenschaftliche Buchgesellschaft.
- Waskan, J. (2008). Knowledge of counterfactual Interventions through cognitive models of mechanisms. *International Studies in Philosophy of Science*, 22 (3), 259-275. [10.1080/02698590802567308](https://doi.org/10.1080/02698590802567308)
- Windt, J. M. (2010). The immersive spatiotemporal hallucination model of dreaming. *Phenomenology and the Cognitive Sciences*, 9 (2), 295-316. [10.1007/s11097-010-9163-1](https://doi.org/10.1007/s11097-010-9163-1)
- Wu, W. (2013). Mental action and the threat of automaticity. In A. Clark, J. Kiverstein & T. Vierkant (Eds.) *Decomposing the will* (pp. 244-261). Oxford, UK: Oxford University Press.

Inference to the Best Prediction

A Reply to Wanja Wiese

Anil K. Seth

Responding to Wanja Wiese's incisive commentary, I first develop the analogy between predictive processing and scientific discovery. Active inference in the Bayesian brain turns out to be well characterized by abduction (inference to the best explanation), rather than by deduction or induction. Furthermore, the emphasis on control highlighted by cybernetics suggests that active inference can be a process of "inference to the best prediction", leading to a distinction between "epistemic" and "instrumental" active inference. Secondly, on the relationship between perceptual presence and objecthood, I recognize a distinction between the "world revealing" presence of phenomenological objecthood, and the experience of "absence of presence" or "phenomenal unreality". Here I propose that world-revealing presence (objecthood) depends on counterfactually rich predictive models that are necessarily hierarchically deep, whereas phenomenal unreality arises when active inference fails to unmix causes "in the world" from those that depend on the perceiver. Finally, I return to control-oriented active inference in the setting of interoception, where cybernetics and predictive processing are most closely connected.

Keywords

Abduction | Control-oriented active inference | Falsification | Objecthood | Presence

Author

Anil K. Seth

a.k.seth@sussex.ac.uk

University of Sussex

Brighton, United Kingdom

Commentator

Wanja Wiese

wawiese@uni-mainz.de

Johannes Gutenberg-Universität

Mainz, Germany

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität

Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University

Melbourne, Australia

1 Introduction

It is a pleasure to respond to [Wanja Wiese's](#) stimulating commentary ([this collection](#)), from which I learned a great deal. Much of what he says stands easily by itself, so here I select just a few key points which warrant further development in light of his analysis.

2 Active inference and hypothesis testing

A central claim in my target paper is that active inference, typically considered as the resolution of sensory prediction errors through action, should also (perhaps primarily) be considered as furnishing disruptive and/or disam-

biguatory evidence for perceptual hypotheses. This claim transparently calls on analogies with hypothesis testing in science (as well as on counterfactually-equipped generative models), and so invites comparisons with theoretical frameworks for scientific discovery, as Wiese nicely develops. In particular, [Wiese](#) notes that I do not "say much about what it takes to disconfirm or falsify a given hypothesis or model", inviting me to "provide a refined treatment of the relation between falsification and active inference" ([this collection](#), p. 2). This is what I shall attempt in this first section.

2.1 The abductive brain

Wiese rightly says that a strict Popperian analogy for active inference is inappropriate since Popperian falsification relies on hypotheses that are derived deductively. Deductive inferences are *necessary inferences*, meaning that their falsification in turn falsifies the premises (theories) from which they derive. Active inference in the Bayesian brain is not deductive for two important reasons. First, as Wiese notes, Bayesian inference is inherently probabilistic so that competing hypotheses become more or less likely, rather than corroborated or falsified. Probabilistic weighting of hypotheses suggests a process of *induction* rather than deduction. Inductive inferences are *non-necessary* (i.e., they are not inevitable consequences of their premises) and are assessed by observation of outcome statistics, by analogy with classical statistical inference. Second, Bayesian reasoning pays attention not just to outcome frequencies but to properties of the explanation (hypothesis) itself, as captured by the slogan that (Bayesian) perception is the brain's "best guess" of the causes of its sensory inputs. This indicates that the Bayesian brain is neither deductive nor inductive but *abductive* (Hohwy 2014), where abduction is typically understood as "inference to the best explanation". In Bayesian inference, what makes a "best" explanation rests not only on outcome frequencies, but also on quantification of model complexity (models with fewer parameters are preferred), and by priors, likelihoods, as well as hyper-priors which may make some prior-likelihood combinations more preferable than others. Importantly, abductive (and inductive) processes are *ampliative*, meaning that they are capable of going beyond that which is logically entailed by their premises. This is important for the Bayesian brain, because the fecundity and complexity of the world (and body) requires a flexible and open-ended means of adaptive response.

So, the Bayesian brain is an abductive brain. But I would like to go further, recalling that active inference enables predictive *control* in addition to perception. This emphasis is particularly clear in the parallels with cybernetics

and applications to interoception developed in the target article, where allostatic¹ control of 'essential variables' is paramount, and where predictive models are recruited towards this goal (Conant & Ashby 1970; Seth 2013). In this light, active inference in the cybernetic Bayesian brain becomes a process of "inference to the best *prediction*", where the "best" predictions are those which enable control and homeostasis under a broad repertoire of perturbations.² It will be interesting to fully develop criteria for "best-making" in this control-oriented form of abductive inference.

2.2 Sophisticated falsificationism, active inference, and model disambiguation

Where does this leave us with respect to theories of scientific discovery? Strict Popperian falsification was already discounted as an analogy for active inference. At the other extreme, parallels with Kuhnian paradigm shifts also seem inappropriate since these are not based on inference whether deductive, inductive, or abductive. Also, such shifts are typically unidirectional: having dispensed with the Copernican worldview once, we are unlikely to return to it in the future. These two points challenge Wiese's analogy between paradigm shifts and perceptual transitions in bistable perception (see Wiese's footnote 12, [this collection](#), p. 9). What best survives in this analogy is an appeal to hierarchical inference, where changes in "paradigm" correspond to alternations between hierarchically deep predictions, each of which recruit more fine-grained predictions which themselves each explain only part of the ongoing sensorimotor flux, under the hyper-prior that perceptual scenes must be self-consistent (Hohwy et al. 2008).

Wiese himself seems to favour Lakatos' interpretation of Popper, a "sophisticated falsificationism" where theories (perceptual hypotheses) can be modified rather than rejected outright, when predictions are not confirmed,

¹ Allostatic: the process of achieving homeostasis.

² There is an interesting analogy here to the overlooked "perceptual control theory" of William T. Powers, which says that living things control their perceived environment by means of their behavior, so that perceptual variables are the targets of control (1973).

and where hypotheses are not tested in isolation (more on this later). As Wiese shows, sophisticated falsification fits well with some aspects of Bayesian inference, like model updating. According to Lakatos, core theoretical commitments can be protected from immediate falsification by introducing “auxiliary hypotheses” which account for otherwise incompatible data (1970). The key criterion - in the philosophy of science sense - is that these auxiliary hypotheses are *progressive* in virtue of making additional testable predictions, as opposed to *degenerate*, which is when the core commitments become less testable.³ This maps neatly to counterfactually-equipped active inference, where hierarchically deep predictive models spawn testable counterfactual sensorimotor predictions which are selected on the basis of precision expectations, and which lead to effective updating (rather than “falsification”) of perceptual hypotheses. As Wiese notes, a good example of this is given by Friston and colleagues’ model of saccadic eye movements (Friston et al. 2012). When it comes to model comparison, sophisticated falsification may even approximate some aspects of abductive inference: “Explaining away is another example of sophisticated falsification. Even when two or more models are compatible with the evidence ... there can be reason to prefer one of them and reject the other” (Wiese this collection, p. 7). This strongly recalls Bayesian model comparison and “inference to the best explanation”, if not its control-oriented “inference to the best prediction” form.

One important clarification is needed about Wiese’s interpretation of model comparison, highlighting the critical roles of action and counterfactual processing. Wiese rightly emphasizes the important insight of Popper and Lakatos that hypotheses are never tested in

isolation, mandating a process of comparison among competing models or hypotheses. However, he implies a sequential testing of each hypothesis: “balloons being launched and then shot down, one by one” (see Wiese this collection, p. 6). This is quite different from the interpretation of model comparison pursued in my target article, where multiple models are considered in parallel, and where counterfactual predictions are leveraged to select the action (or experiment) most likely to *disambiguate* competing models. In Bayesian terms this is reflected in a shift towards model comparison and averaging (FitzGerald et al. 2014; Rosa et al. 2012), as compared to inference and learning on a single model. Bongard and colleagues’ evolutionary robotics example was selected precisely because it illustrates this point so well (Bongard et al. 2006). Here, repeated cycles of model selection and refinement lead to the prescription of novel actions that best disambiguate the current best models (note the plural). Indeed, it is the repeated refinement of disambiguatory actions that gives Bongard’s starfish robot its compelling “motor babbling” appearance. To reiterate: different actions may be specified when the objective is to disambiguate multiple models in parallel, as compared to testing models one-at-a-time. In the setting of the cybernetic Bayesian brain this example is important for two reasons: it underlines the importance of counterfactual processing (to drive the selection of disambiguatory actions) and it emphasizes that predictive modelling can be seen as a means of control in addition to discovery, explanation, or representation. In this sense it doesn’t matter how accurate the starfish self model is – what matters is whether it works.

2.3 Science as control or science as discovery?

The distinction between explanation and control returns us to the philosophy of science. Put simply, the views of Popper, Lakatos, and (less so) Kuhn, are concerned with how science reveals truths about the world, and how falsification of testable predictions participates in this process. Picking up the threads of abduction,

³ An important application of this idea is to the Bayesian brain itself as a scientific hypothesis. A concern about the Bayesian brain hypothesis is that it can be insulated from falsification by postulating convenient (typically unobservable) priors, much like adaptationist explanations in evolutionary biology can be critiqued as “just so” stories. The key question, not answered here, is whether neural mechanisms implement (approximations to) Bayesian inference, or whether Bayesian concepts merely provide a useful interpretative framework. In the former case one would require the Bayesian brain hypothesis to be progressive not degenerate.

control-oriented active inference, and “inference to the best prediction”, we encounter the possibility that theories of scientific discovery might themselves appear differently when considered from the perspective of control. Historically, it is easy to see the narrative of science as a struggle to gain increasing control over the environment (and over people), rather than a process guided by the lights of increasing knowledge and understanding.⁴ A proper exploration of this territory moves well beyond the present scope (see e.g., Glazebrook 2013). In any case, whether or not this perspective helps elucidate scientific practice, it certainly suggests important limits in how far analogies can be taken between philosophies of scientific discovery and the cybernetic Bayesian brain.

3 Perceptual presence and counterfactual richness

The second part of Wiese’s commentary picks up on the issue of *perceptual presence*, which in my target article was associated with the “richness” of counterfactual sensorimotor predictions (see also Seth 2014, 2015b). Wiese makes a number of connected points. First, he rightly notes an ambiguity between objecthood and presence in perceptual phenomenology, as presented in my target article (Seth this collection) and in Seth (2014). Second, he introduces the notion of *causal encapsulation* as a third phenomenological dimension, complementing counterfactual richness and perspective dependence. He spends some time developing examples based on cognitive phenomenology and mental action to illustrate how these dimensions might relate. Here, I will focus on the relationship between presence and objecthood from the perspective of counterfactual predictive processing – or more specifically the theory of “Predictive Processing of SensoriMotor Contingencies” (PPSMC; Seth 2014, 2015b).⁵

⁴ The continually increasing pressure to justify research in terms of “impact” – especially when seeking funding – highlights one way in which an emphasis on control (rather than discovery) is realized in scientific practice.

⁵ See also my response (Seth 2015b) to commentaries on (Seth 2014), which focuses on this issue.

3.1 Presence and objecthood together

As Wiese notes, when visually perceiving a real tomato (figure 1A) there is both a sense of *presence* (the subjective sense of reality of the tomato) and of *objecthood* (the perception that a (real) object is the cause of sensations). Importantly, while distinct, these properties are not independent. There is a “world-revealing” dimension to perceptual presence which is closely aligned with the experience of an externally-existing object: “How can it be true ... that we are perceptually aware, when we look at a tomato, of the parts of the tomato which, strictly speaking, we do not perceive. This is the puzzle of perceptual presence” (Noë 2006, p. 414).

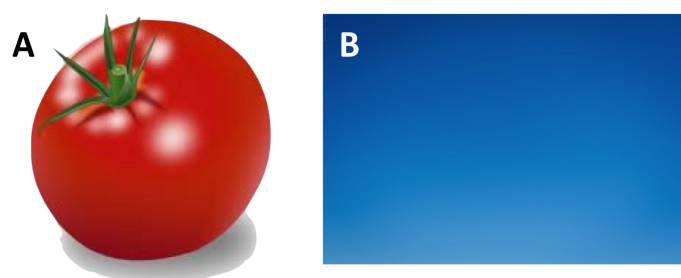


Figure 1: A. An image of a tomato. B. An image of a clear blue sky.

How does this object-related world-revealing presence come about? In predictive processing (and by extension PPSMC), objecthood depends on predictive models encoding hierarchically deep invariances that accommodate complex nonlinear mappings from (object-related, world-revealing) hidden causes to sensory signals (Clark 2013; Hohwy 2013). There is a reciprocal dependency here between hierarchical depth and counterfactual richness, because (i) hierarchically deep invariances in generative models enable precise predictions about rich repertoires of counterfactual sensorimotor mappings, and (ii) counterfactual richness can scaffold the acquisition of hierarchically deep invariant predictions. One might even say that hierarchically deep invariances are partly *constituted* by (possibly latent) predictions of counterfactually rich sensorimotor mappings (Seth 2015b). These dependencies indicate that ob-

jecthood and world-revealing presence depend on *expectations about counterfactual richness*, rather than counterfactual richness *per se*. Altogether, counterfactually-informed active inference enables the extraction and encoding of hierarchically deep hidden causes of sensory signals. In virtue of hierarchical depth, these inferred causes will also be *perspective invariant*, in the sense that they will have been separated from those causes that depend on on actions (or other properties) of the perceiver (see [Wiese this collection](#), p. 11). In short, to the extent that objecthood and perceptual presence go together, so do hierarchical depth (encoding world-revealing invariances) and (expected) counterfactual richness.

3.2 Presence and objecthood apart

So far so good, but it is evident that presence and objecthood do not *always* go together ([Di Paolo 2014](#); [Froese 2014](#); [Madary 2014](#)), a phenomenological fact which requires further analysis ([Seth 2015b](#)). Presence without objecthood is exemplified in vision by the experience of a uniform deep blue sky (Figure 1B), and is also characteristic of non-visual modalities like olfaction ([Madary 2014](#)). The visual impression of a blue sky, or the tang of briny sea air, both seem perceptually present but without eliciting any specific phenomenology of objecthood. At the same time, the corresponding predictive models are likely to be hierarchically shallow and counterfactually poor: there is not much I can do (besides closing my eyes or looking away) to alter the sensory input evoking a blue-sky experience, and the inferred hidden causes are unlikely to lie behind multiple inferential layers. Hierarchical shallowness may explain the lack of phenomenal objecthood, but why isn't there also a lack of perceptual presence?

Blue-sky-experiences (and olfactory scenes) actually *do* lack the world-revealing presence associated with objecthood. But they do not appear *phenomenally unreal* in the sense that perceptual afterimages and synaesthetic concurrents are experienced as unreal. In PPSMC, phenomenal unreality can arise from an inferential failure to separate hidden causes

in the world, from those that depend on actions (or other properties) of the perceiver ([Seth 2015b](#)). This in turn emerges from violations of counterfactual predictions. For example, consider how saccadic eye movements engage counterfactual predictions. Perceptual afterimages track eye movements, violating counterfactual predictions associated with world-revealing hidden causes that rest on active inference. In contrast, counterfactual predictions associated with blue skies are less amenable to disconfirmation by eye movements, so (non-object-related) perceptual presence remains.⁶

Summarizing, perceptual presence, as an explanatory target, can be refined into (i) a *world-revealing presence* associated with objecthood and hierarchical depth, and (ii) a *phenomenal unreality* arising from a failure to inferentially separate hidden causes in the world from those associated with the perceiver. Both rely on counterfactual processing, and so both call on active inference. Perspective invariance is also implicated in objecthood (through hierarchical depth) and phenomenal unreality (through isolating worldly causes), suggesting that this dimension may not be as separable from counterfactual richness as proposed by [Wiese \(this collection, p. 13\)](#). But is that all there is to presence?

3.3 Causal encapsulation and embodiment

Wiese distinguishes three dimensions to perceptual presence: counterfactual richness (vs. poverty), perspective invariance (vs. dependence), and causal encapsulation (vs. integration). The third of these, causal encapsulation, is perhaps the hardest to pin down. The idea as I understand it, is that a representation (predictive model) is causally encapsulated if it is inferentially isolated from other hidden causes; by contrast it is causally *open* or *integrated* if it expresses a rich set of relations to other inferred

⁶ Phenomenal unreality on this story corresponds to a loss of “transparency” as described by ([Metzinger 2003](#)). For Metzinger, transparency is lost – and phenomenal unrealness results – when the “construction process” underlying perception becomes available for attentional processing. This maps neatly on a failure to inferentially unmix world-related from perceiver-related hidden causes – see [Seth \(2015b\)](#) for more on this.

causes. So, a predictive model underlying the experience of a tomato may be causally integrated with that underlying the experience of the table on which it lies, and the hand (maybe my hand), which is poised to reach out and pick it up. Here, there may be a relation between causal encapsulation/integration and the inferential unmixing of perceiver-related and world-related hidden causes: a failure to separate these causes would presumably prevent rich causal integration with other hidden causes in the world.

The concept of causal encapsulation highlights another interesting aspect of Wiese's commentary: the idea that counterfactual predictions may not always encode sensorimotor contingencies: "it might be equally relevant to encode how sensory signals pertaining to the tomato would change if the wind were to blow ... or if the tomato were to fall down" (Wiese [this collection](#), p. 11). While such extra-personal causal contingencies may be salient in many cases, I see them as secondary to sensorimotor body-related counterfactual predictions. By definition they do not involve active inference: I have to wait for the wind to change direction (though perhaps I might move to get a better view). This means that many central features of active inference discussed here – its relation to predictive control, homeostasis, and counterfactually-informed model disambiguation – do not apply.

The body re-emerges here as central, this time as a ground for the generation of counterfactual predictions. Specifically, bodily constraints shape counterfactual predictions since they place limits on how actions can be deployed in intervening upon the (inferred) causes of sensory input. This suggests that changing action repertoires would alter experiences of presence. Wiese raises out-of-body-experiences and dream experiences as a relevant context ([this collection](#), p. 15), where subjects sometimes identify their first-person-perspective, not with a body, but with an unextended point in space. I agree with him that examining world-revealing presence in these situations would be fascinating, if extremely difficult in practice.

The body is of course not only a source of counterfactual predictions, but also the target of counterfactually-informed active inference, both for representation (exemplified by the rubber-hand-illusion, as mentioned by Wiese) and for control.⁷ As emphasized in the target article, control-oriented active inference is particularly significant for *interoception*, where predictive modelling is geared towards allostasis and homeostasis rather than accurate representation (see also [Seth 2013](#)). Returning the focus to interoceptive inference raises a host of intriguing questions, which can only be gestured at here. One may straightaway wonder how counterfactual aspects of interoceptive inference shape the "presence" of emotional and body-related experiences. Is it possible to have an emotional experience lacking in "affective presence" – and what is the phenomenological correlate of "objecthood" for interoceptive experience? Other interesting questions are how precision weighting sets the balance between representation versus control in active interoceptive inference, and what it means to isolate "wordly" causes when both the means and the targets of active inference are realized in the body. These are not just theoretical questions: advances in virtual reality ([Suzuki et al. 2013](#)) and in methods for measuring interoceptive signals ([Hallin & Wu 1998](#)) promise real empirical progress on these issues.

4 Conclusions

This response has been shaped by Wiese's perspicuous focus on the philosophy of science and on the phenomenology of perceptual presence. My response to the first topic was to frame the Bayesian brain in terms of *control-oriented ab-*

⁷ Wiese, when discussing König's FeelSpace project ([Kaspar 2014](#)), interprets PPSMC as saying that increased practice with the FeelSpace compass belt – and hence increased counterfactual richness – would lead to "increased perceptual presence (for the belt, or the vibrations, or the hip/waist, etc.)" (Wiese [this collection](#), p. 17). I see things differently. The counterfactual predictions, while mediated by the belt, relate to hidden causes in the world (e.g., magnetic north). In fact, PPSMC says that FeelSpace practice would lead to hierarchically deep and counterfactually rich models of how "magnetic north" impacts on belt vibrations and the like, leading to increased world-revealing presence for these worldly causes but diminished perceptual presence of the tactile stimulation itself. Still, the FeelSpace project certainly provides a fertile empirical testbed for the ideas raised here.

duction, where falsification is replaced by “inference to the best prediction” as a criterion for progress. I also reinforced the dependency between active inference and counterfactual processing, which underpins the important case of disambiguatory active inference in Bayesian model comparison. With respect to perceptual presence I proposed a distinction between world-revealing presence and phenomenal unreality (Seth 2015b). World-revealing presence corresponds to objecthood and is associated with hierarchical depth, expected counterfactual richness, and perspective invariance of perceptual hypotheses. Phenomenal unreality transpires when perceptual inference fails to unmix world-related from perceiver-related causes; this corresponds to a loss of “phenomenal transparency” (Metzinger 2003) and depends on violation of counterfactual sensorimotor predictions. Space constraints prevented me considering Wiese’s discussion of the “presence” of cognitive phenomenology, like abstract mathematical and philosophical thinking, in these terms. There is of course a rich literature in linking such phenomena to the body (Lakoff & Nunez 2001), and hence perhaps to active inference where the concept of a “mental action” becomes critical (O’Brien & Soteriou 2009). Space constraints also prevented Wiese from elaborating on interoception, which I consider the most interesting setting for control-oriented active inference, in virtue of the cybernetics-inspired emphasis on homeostasis and allostasis. Interesting questions emerge here about how counterfactual processing plays into the phenomenology of interoceptive experience.

Cognitive scientists have long argued for a continuity between perception and action (Dewey 1896). To close, I suggest thinking instead of a continuum between *epistemic* and *instrumental* active inference. This is simply the idea that active inference – a continuous process involving both perception and action – can be deployed with an emphasis on predictive control (instrumental), or on revealing the causes of sensory signals (epistemic). This process intertwines interoception, proprioception, and exteroception, and autonomic and motoric action, with the balance always delicately orchestrated

by precision optimisation and counterfactual processing. Putting things this way provides a new way to link “life” and “mind” (Godfrey-Smith 1996) and may help reveal the biological imperatives underlying perception, emotion, and selfhood.

Acknowledgements

I am grateful to the Dr. Mortimer and Theresa Sackler Foundation, which support the work of the Sackler Centre for Consciousness Science. Many thanks to Thomas Metzinger, Jennifer Windt and the MIND group for inviting me to participate in this project, to Jakob Hohwy and Karl Friston for correspondence about abductive inference, and to Wanja Wiese for his excellent commentary.

References

- Bongard, J., Zykov, V. & Lipson, H. (2006). Resilient machines through continuous self-modeling. *Science*, 314 (5802), 1118-1121. [10.1126/science.1133687](https://doi.org/10.1126/science.1133687)
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioural and Brain Sciences*, 36 (3), 181-204. [10.1017/S0140525X12000477](https://doi.org/10.1017/S0140525X12000477)
- Conant, R. & Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 1 (2), 89-97.
- Dewey, J. (1896). The reflex arc concept in psychology. *Psychological Review*, 3, 357-370.
- Di Paolo, E. A. (2014). The worldly constituents of perceptual presence. *Frontiers in Psychology*, 5. [10.3389/fpsyg.2014.00450](https://doi.org/10.3389/fpsyg.2014.00450)
- FitzGerald, T. H., Dolan, R. J. & Friston, K. J. (2014). Model averaging, optimal inference, and habit formation. *Frontiers in Human Neuroscience*, 8. [10.3389/fnhum.2014.00457](https://doi.org/10.3389/fnhum.2014.00457)
- Friston, K. J., Adams, R. A., Perrinet, L. & Breakspear, M. (2012). Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology*, 3. [10.3389/fpsyg.2012.00151](https://doi.org/10.3389/fpsyg.2012.00151)
- Froese, T. (2014). Steps toward an enactive account of synesthesia. *Cognitive Neuroscience*, 5 (2), 126-127. [10.1080/17588928.2014.905521](https://doi.org/10.1080/17588928.2014.905521)
- Glazebrook, T. (Ed.) (2013). *Heidegger on science*. New York, NY: State University of New York Press.
- Godfrey-Smith, P. G. (1996). Spencer and Dewey on life and mind. In M. Boden (Ed.) *The philosophy of artificial life* (pp. 314-331). Oxford, UK: Oxford University Press.
- Hallin, R. G. & Wu, G. (1998). Protocol for microneurography with concentric needle electrodes. *Brain Research Protocols*, 2 (2), 120-132.
- Hohwy, J. (2013). *The predictive mind*. Oxford, UK: Oxford University Press.
- (2014). The self-evidencing brain. *Noûs*. [10.1111/nous.12062](https://doi.org/10.1111/nous.12062)
- Hohwy, J., Roepstorff, A. & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, 108 (3), 687-701. [10.1016/j.cognition.2008.05.010](https://doi.org/10.1016/j.cognition.2008.05.010)
- Kaspar, K., König, S., Schwandt, J. & König, P. (2014). The experience of new sensorimotor contingencies by sensory augmentation. *Consciousness and Cognition*, 28. [10.1016/j.concog.2014.06.006](https://doi.org/10.1016/j.concog.2014.06.006)
- Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.) *Criticism and the growth of knowledge* (pp. 91-196). Cambridge, UK: Cambridge University Press.
- Lakoff, G. & Nunez, R. (2001). *Where mathematics comes from: How the embodied mind brings mathematics into being*. New York, NY: Basic Books.
- Madary, M. (2014). Perceptual presence without counterfactual richness. *Cognitive Neuroscience*, 5 (2), 131-133. [10.1080/17588928.2014.907257](https://doi.org/10.1080/17588928.2014.907257)
- Metzinger, T. (2003). Phenomenal transparency and cognitive self-reference. *Phenomenology and the Cognitive Sciences*, 2, 353-393.
- Noë, A. (2006). Experience without the head. In T. Gendler & A. Hawthorne (Eds.) *Perceptual experience* (pp. 411-434). New York, NY: Clarendon / Oxford University Press.
- O'Brien, L. & Soteriou, M. (Eds.) (2009). *Mental actions*. Oxford, UK: Oxford University Press.
- Powers, W. T. (1973). *Behavior: The control of perception*. Hawthorne, NY: Aldine de Gruyter.
- Rosa, M. J., Friston, K. J. & Penny, W. (2012). Post-hoc selection of dynamic causal models. *Journal of Neuroscience Methods*, 208 (1), 66-78. [10.1016/j.jneumeth.2012.04.013](https://doi.org/10.1016/j.jneumeth.2012.04.013)
- Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, 17 (11), 565-573. [10.1016/j.tics.2013.09.007](https://doi.org/10.1016/j.tics.2013.09.007)
- (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive Neuroscience*, 5 (2), 97-118. [10.1080/17588928.2013.877880](https://doi.org/10.1080/17588928.2013.877880)
- (2015). The cybernetic bayesian brain. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- (2015b). Presence, objecthood, and the phenomenology of predictive perception. *Cognitive Neuroscience*
- Suzuki, K., Garfinkel, S. N., Critchley, H. D. & Seth, A. K. (2013). Multisensory integration across exteroceptive and interoceptive domains modulates self-experience in the rubber-hand illusion. *Neuropsychologia*, 51 (13), 2909-2917. [10.1016/j.neuropsychologia.2013.08.014](https://doi.org/10.1016/j.neuropsychologia.2013.08.014)
- Wiese, W. (2015). Perceptual presence in the Kuhnian-Popperian Bayesian brain. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.

The Ongoing Search for the Neuronal Correlate of Consciousness

Wolf Singer

A few decades ago the search for the neuronal correlates of consciousness was considered both technically intractable and philosophically questionable. Searching for a material substrate of phenomena accessible only from the first-person perspective appeared to be epistemically problematic. But the development of non-invasive imaging technologies and the availability of intracranial recordings from patients alleviated the imminent technical problems. Progress in the analysis of the connectome of the brain, and the introduction of multisite recordings from the cerebral cortex of animals led to a revision of concepts in the field of cognitive neuroscience, emphasizing principles of distributed processing in recurrent networks with non-linear dynamics, self-organization, and coding in high-dimensional-state space. These advances, together with the growing evidence for epigenetic shaping of brain functions by socio-cultural influences, pave the way for novel theories that attempt to bridge the gap between neuronal processes and subjective states.

Keywords

Binding problem | Cultural evolution | Distributed processing | Epigenetic shaping | Long-range synchronisation of oscillation | Meta-representation | Naturalistic epistemology | Neural correlate of consciousness (NCC) | Oscillations | Perceptual constancy | Small-world architecture | Subconscious processing | Synchrony | Unity of consciousness | Workspace of consciousness

Author

Wolf Singer

w.singer@brain.mpg.de

Max Planck Institute for Brain Research (MPI)
Frankfurt a. M., Germany

Commentator

Valdas Noreika

valdas.noreika@mrc-cbu.cam.ac.uk

Medical Research Council
Cambridge, United Kingdom

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

Progress in brain research, especially in the domain of cognitive neuroscience, renders phenomena that have traditionally been subjects of the humanities amenable to scientific investigation. It has now become possible to investigate the neuronal underpinnings of mental phenomena such as perception, decision making, control of attention, language perception and production, action planning, storage and recall of memories, emotions and moods, desires and aversions, and last but not least consciousness. This research agenda is confronted with a number of fascinating challenges. One is the immense complexity

of the brain processes that underlie these highly-differentiated cognitive functions. Another results from the fact that many of the phenomena whose neuronal correlates are to be investigated are subjective phenomena, accessible and describable only from a first-person perspective. Hence there is an epistemically problematic gap between what is observable from the third-person perspective of scientific inquiry and the explananda that need to be defined in terms of first-person experience. Yet another challenge is that a relatively young scientific discipline is set to enter territories that

for millennia have been ploughed by great minds who have coined terms, formulated concepts, and constructed belief systems based on evidence extracted from introspection, intuition, and observations that relied exclusively on the natural senses. This raises numerous and on occasions frustrating problems for communication, because bridges have yet to be built between the more recent naturalistic description systems and the highly-differentiated terminology nurtured in the humanities. Some of these problems surface in passionate discussions on the existence of free will, the nature of perception, the constitution of the Self, and intentionality and mental causation—and above all on the question of whether it is even possible to identify neuronal correlates of mental, subjective phenomena.

In this chapter some of these challenges will be discussed from a neurobiological perspective. We shall first review the state of the art in the field of cognitive neuroscience, emphasizing recent changes in our views on the brain. These have been forced upon us by the novel data produced by new and powerful technologies. These insights show the brain to be a highly distributed, self-active system with non-linear dynamics; rather than a hierarchically-organized stimulus-response machine, as has been proposed by behaviourist theories. Subsequently, an excursion will be made into epistemology, to establish the extent to which brain research can contribute to philosophical discussions concerning the nature of perception. The process of perceiving will be interpreted as a constructive act in which sparse sensory signals are matched with a huge amount of stored knowledge; and the various sources of this knowledge will be discussed. This section will set the stage for the following discussion of the putative neuronal correlates of consciousness (NCC), as it will highlight the cognitive constraints and idiosyncrasies of cognitive systems that owe their abilities to evolutionary processes. Before reviewing various theories on the NCC, an attempt will be made to define the explanandum—in full awareness of the futility of this attempt. The review of experimental work in search of the NCC will be followed by a brief

account of our own experimental contributions, and then an attempt will be made to demystify the so-called “hard problem” of consciousness—the problem of finding a naturalistic explanation for the qualia, namely these immaterial connotations of our experiences. My proposal will be that the problem can be alleviated if we consider not only individual brains and biological evolution but also cultural evolution and the social realities that have emerged from socio-cultural interaction between human beings.

As this contribution addresses an interdisciplinary audience I considered it appropriate to not only refer to published work when alluding to experimental findings and concepts but to sometimes provide explicit and detailed background information. To this end I have adapted passages from a few of my own publications, in which I had addressed some of the issues that needed to be recapitulated for the sake of clarity in the present contribution. These passages include descriptions of experimental findings from my lab and descriptions of the state of the art in the neurosciences.

2 The state of the art in cognitive neuroscience: A paradigm shift

2.1 Classical views

As detailed in several previous reviews the neurosciences are about to undergo a paradigm shift towards concepts that consider the brain as a self-organizing complex system with non-linear dynamics that exploits a huge body of stored knowledge to interpret sensory signals, formulate hypotheses and to generate predictive models of the world in order to optimize adapted behavioral responses (Singer 2009, 2013). For many decades, the search for the neuronal underpinnings of cognitive and executive functions has been guided by the behaviourist view that the brain is essentially a highly complex and versatile stimulus-response machine, in which serial processing strategies prevail. This view received further support from early anatomical data that emphasized that feed-forward connections exhibit high topographic precision and possess strong driving synapses, while feedback

connections are diffuse and only modulatory. The rather impressive performance of artificial pattern-recognition systems based on such processing architectures suggested that neuroscientists were on the correct path. Accordingly, they set out to study the responses of neurons to sensory stimulations across the various stages of the processing hierarchy and analyzed activation patterns associated with motor output, hoping that these strategies would eventually lead to reductionist explanations of the neuronal mechanisms that support cognition, memory, decision making, planning, and motor behaviour. The strategy to follow the transformation of activity from the sensory surfaces over the numerous levels of hierarchically-organized processing structures to the respective effector organs proved to be extremely fruitful. Comparison of brains from different species provided compelling evidence that the basic principles according to which neurons function and exchange signals have been preserved throughout evolution with only minor modifications. For the comparatively simple nervous systems of certain invertebrates, this behaviourist approach allowed for near-complete descriptions of the neuronal mechanisms underlying particular behavioural manifestations. This nurtured the expectation that pursuing this research strategy would sooner or later allow us to explain in the same way the more complex behaviour of mammals—and ultimately also the highly-differentiated cognitive functions of primates and human subjects. However, in recent decades the pursuit of this approach has led to an accumulation of evidence that demands a revision of the classical hypothesis, which emphasizes serial feed-forward processing of sensory information within hierarchically-organized architectures.

2.2 Observations forcing an extension of classical views

Advances in the analysis of the cortical connectome, the introduction of multisite recording techniques, and the development of imaging methods assessing whole-brain activity have generated data that necessitate an extension of

classical views, raise novel questions, and likely provide new solutions to old problems.

Anatomical evidence: i) Within processing streams from sensory surfaces to executive organs, feedback projections are in general more numerous than feed-forward projections, emphasizing the importance of top-down control (Felleman & van Essen 1991). ii) Connections linking neurons within distinct cortical areas cross the boundaries between areas (Schwarz & Bolz 1991). Thus, the cerebral cortex appears to be a continuously coupled sheet, the different cortical areas being distinguished mainly by their input and output connections. iii) From primary sensory areas onwards, processing streams diverge into numerous parallel pathways whose nodes are linked by massive reciprocal connections, both within and across modalities (Markov & Kennedy 2013; Markov et al. 2013). iv) The rule that feed-forward connections originate in supra- and feedback projections in infragranular layers does not hold for nearby cortical areas (Markov & Kennedy 2013; Markov et al. 2013). Together with electrophysiological evidence (De Pasquale & Sherman 2011), this threatens the strict distinction between feed-forward driving and feedback modulatory connections. v) Finally, statistical analysis of interareal connectivity suggests an organisation resembling small-world, rich-club networks (see Van den Heuvel & Sporns 2013) that minimize path length between nodes (areas; Van den Heuvel & Sporns 2011; Sporns 2013). However, analysis of projections with cellular resolution suggests as one reason for short path length the surprisingly high degree of connectedness among cortical areas. Statistical analysis suggests that more than 60% of possible links between network nodes are actually realized (Markov & Kennedy 2013).

Functional evidence: i) Even in early sensory areas neurons lose their simple feature-specific responses when challenged with complex stimuli (David et al. 2004; Vinje & Gallant 2000). Moreover, responses are influenced by stimuli in other modalities, by attention, reward expectation, and contents in working memory, thus suggesting contextual modulation not only by intrinsic connections but also by top-down

projections (Engel et al. 2001; Calvert et al. 1997; Iurilli et al. 2012; Muckli & Petro 2013; Stokes et al. 2013). ii) The notion of strictly serial processing from input layer four, via layers three and two, to the output layers five and six of the cerebral cortex needs to be revised in light of evidence that vigorous responses can also be elicited by sensory input when parts of this canonical circuit are disrupted (Constantinople & Bruno 2013). The possibility that supra and infragranular compartments can operate in parallel is further supported by evidence that the two subdivisions engage in oscillatory activity in different frequency bands (gamma in supra- and alpha or beta in infragranular layers; Buffalo et al. 2011; Roopun et al. 2008). iii) Multisite recordings indicate that “spontaneous” fluctuations in the responsiveness of individual neurons are often the reflection of coordinated, highly structured spatio-temporal activity patterns rather than the result of noise (Kenet et al. 2003; Fries et al. 2001a). iv) Widely distributed cortical areas exhibit coherent fluctuations of their spontaneous activity, forming functionally-coupled networks that change in their composition in a state-dependent way (Fox et al. 2005; Hipp et al. 2012; Raichle 2011; Raichle et al. 2001). Thus, the cortex—and in a wider sense the brain—appears to be a highly active, pattern-generating system, rather than just a stimulus-driven device. v) Analysis of whole brain activity with functional magnetic resonance imaging (fMRI) and electroencephalographic (EEG) and magnetoencephalographic (MEG) measurements indicates that virtually all cognitive and executive functions are associated with the activation of networks of often widely-distributed cortical areas (Engen & Singer 2013; Friederici & Gierhan 2013; Hipp et al. 2011; Hodzic et al. 2009; Power & Petersen 2013). This suggests that distributed networks are a substrate of functions rather than individual specialized structures. vi) Finally, analysis of the brain’s dynamic signatures indicates that neuronal populations can engage in oscillatory activity in characteristic frequency bands and synchronize their discharges, such that the respective frequency bands and the composition of coherently active

cell groups depend on central states, attention, cognitive tasks, and goals of action (Buzsáki 2006; Singer 2010).

These novel anatomical and functional data suggest as a prevailing organizational principle distributed processing in densely coupled, recurrent networks with non-linear dynamics, which are capable of supporting high dimensional states. This organization requires a high degree of coordination of distributed processes, suggesting that special mechanisms are implemented to dynamically bind local processes into coherent global states, and to configure functional networks “on the fly” in a context- and goal-dependent way. It has also become clear that the brain is by no means a stimulus-driven system. Rather, it is self-active, permanently generating highly structured, high-dimensional spatio-temporal activity patterns. These patterns are far from being random, and instead seem to reflect the specificities of the functional architecture that is determined by genes, modified by experience throughout post-natal development, and further shaped by learning. These self-generated activity patterns in turn seem to serve as priors with which incoming sensory signals are compared. Perception is now understood as an active, reconstructive process, in which self-generated expectancies are compared with incoming sensory signals. The development of methods that allow simultaneous registration of the activity of large numbers of spatially-distributed neurons revealed a mind-boggling complexity of interaction dynamics—which in turn eludes the capacity of conventional analytical tools and, because of its non-linearity, challenges hypotheses derived from intuition.

In the last decade theoreticians have begun to explore and appreciate the immense computational power of such self-organizing recurrent networks that gave rise to concepts such as “reservoir computing”, “echo-state computing” or “liquid computing” (Buonomano & Maass 2009; Lukoševičius & Jaeger 2009). The evidence that resting-state activity is highly structured, that information is contained in the spatio-temporal relations between the responses of widely distributed neurons, and that stimulus-response functions depend crucially on state

variables generated within the brain are in principle compatible with such advanced concepts of information processing in highly non-linear, high-dimensional dynamic systems; but neurobiological approaches taking such considerations into account are still very rare.

2.3 Persisting explanatory gaps

Our rather detailed knowledge of the response properties of individual neurons in different brain structures, and of the microcircuits that shape these responses, stands in stark contrast to our ignorance of the complex and highly dynamic processes through which the myriads of spatially-distributed neurons interact in order to produce specific behaviours. Evidence from invasive and non-invasive multi-site recordings indicates that most higher brain functions result from the coordinated interaction of large numbers of neurons, which become associated in a context- and goal-dependent way into ad-hoc formed functional networks. These networks are dynamically configured on the backbone of the anatomical connections (for review see [von der Malsburg et al. 2010](#)). Evidence also indicates that these interactions give rise to extremely complex spatio-temporal patterns that are characterized by oscillations in a large number of different frequency bands, which can synchronize, exhibit phase shifts, and even cross frequency coupling ([Uhlhaas et al. 2009](#)). In the light of these novel data, the brain—and in particular the neocortex—appears to be a self-active, self-organizing “complex system” which exhibits non-linear dynamics, is capable of utilizing multiple dimensions for coding (space, amplitude, oscillation frequency, and phase), operates in a tightly-controlled range of self-organized criticality ([Shew et al. 2009](#); edge of chaos), and constantly generates highly-structured, high-dimensional activity patterns that are likely to represent stored information. However, how exactly information is encoded in the trajectories of these high-dimensional and non-stationary time series is largely unknown, and is the subject of increasingly intense research. Moreover, with the exception of a few studies in which selective manipulation of the activity of

defined neuron groups were shown to affect behaviour in a particular way ([Salzman et al. 1992](#); [Houweling & Brecht 2008](#); [Han et al. 2011](#)) most of the available evidence on the relations between neuronal responses and behaviour is still correlative in nature. This makes it difficult to determine whether an observed variable is an epiphenomenon of a hidden underlying process or is causally involved in accomplishing a particular function. Thus, systems neuroscience now faces the tremendous challenge of analyzing the principles of distributed dynamic coding and of obtaining causal evidence for the functional role of specific activation patterns, in order to distinguish between functionally-relevant variables and epiphenomena.

In conclusion, we have to abandon classical notions of the neuronal representation of perceptual objects and, in the same vein, that of motor commands. The consequence is that it becomes once again unclear how the distributed processes that deal with the various properties of a perceptual object—its visual, haptic, acoustic, olfactory and gustatory features—are bound together in order to give rise to a coherent representation or percept. Given this, it may appear more than bold to attempt to identify the neuronal correlates of consciousness—probably the highest and most mysterious of our cognitive functions.

2.4 What neuroscientists believe

Despite the numerous gaps in our understanding of integrated brain functions, neurobiologists agree on a number of general conclusions on the relation between brain processes and behavioural phenomena. The majority of neurobiologists seem to consent that all cognitive and executive functions that we can observe in human beings, including the highest mental activities and consciousness, are the result, not the cause, of neural interactions. Consequently, it is held that mental phenomena follow or emerge from neural interactions and do not precede them. Furthermore, it is assumed that all neural processes obey the known laws of nature. The reason for this is that the behaviour of organisms of low complexity, such as, for example,

molluscs or worms, can be fully explained by registering the activity of their neurons and establishing causal relations between the spatio-temporal patterns of this activity and the respective behaviour. There is, at present, no need to postulate any additional unknown forces, laws, or modes of interaction in order to explain their behaviour. The reason for this is that evolution is a very conservative process. Once an invention has been made that increases fitness it tends to be conserved, unless there is a major change in conditions that makes this invention obsolete or maladapted. Therefore, our nerve cells function in exactly the same way as those of snails. Likewise, the development of structures also follows a very conservative path. Since the first appearance of the cerebral cortex, the six-layered sheet of nerve cells that covers the hemispheres of the brain, no new structures have emerged. There is just more of the same, and this increase in complexity marks the difference between the brain of a human being and that of our nearest neighbours, the great apes. Apparently, this processing substrate and the associated gain of complexity marks the difference between species that failed and those that succeeded in promoting cultural evolution—with all its far reaching consequences. In this context, however, one needs to consider that cultural evolution created a socio-cultural environment of ever-increasing complexity that in turn contributes to the epigenetic shaping of brain architectures. Thus, even if the genetically-determined layout of brain architectures has changed little since the beginning of human civilisation, those features that can be modified by epigenetic shaping are likely to have undergone major modifications. This fact has not always been taken into account in the past; but its implications will be discussed below. But this additional twist concerns the epigenetic modifiability of our brains, and not its basic functional principles.

3 Contributions of neuroscience to philosophy

Once the neurosciences began to investigate the neuronal underpinnings of higher cognitive func-

tions, especially those realized in human brains, an increasing number of questions, traditionally investigated by the humanities, were addressed through empirical studies within the rapidly developing field of cognitive neuroscience. One obvious domain for this investigation was epistemology. Cognitive neuroscience explores from a third-person perspective the mechanisms that mediate our perception and the acquisition of knowledge. Longstanding discussions about the objectivity of cognition, the question of how constructive our perceptual processes really are, and how reliable or idiosyncratic they might be, need to be reconsidered on the basis of neurobiological data. Another question, to which the neurosciences will have to find an answer, is related to the mind-body problem: how can mental phenomena, namely immaterial entities such as the qualia of perception and social realities such as belief and value systems, emerge from the material interactions between nerve cells in human brains? These immaterial phenomena came into this world once the cognitive abilities of *Homo sapiens* initiated the evolution of cultures. They affect our lives as much as the material constraints of the world in which we evolve, but they have a different ontological status to the neuronal processes that brought them into this world. Yet another question that solicits discussion between neuroscientists and philosophers of mind is the nature of consciousness. The question of the constitution of the intentional Self is closely related to this issue, as is the conundrum of the existence of Free Will. If the material processes in individual brains and the social realities resulting from the interactions of humans are the basis and cause of mental phenomena, and if brain processes follow the known laws of nature, then there ought to be unifying description systems that bridge the gap between phenomena assessed from third- and first-person perspectives. If such approaches turn out to be feasible philosophical positions, the postulation of an ontological dualism will have to be modified. This will have far-reaching consequences for our self-understanding and the delineation of the border between “physics” and “metaphysics”.

3.1 An epistemic caveat

It is obvious that our perceptions and imaginations, as well as our ability to reason, are constrained by the cognitive abilities of our brains—and brains, like all other organs, are the product of an evolutionary process. Hence our brains have become adapted to the conditions of the mesoscopic world in which life has evolved. This is the world within the scale of millimeters to meters, it is the world where the laws of classical physics prevail; it is not the world of quantum physics and it is not the world of astrophysics. As a consequence, our cognitive functions have become adjusted to assure survival in this mesoscopic world. Problem-solving in this dangerous and poorly-predictable world requires the application of pragmatic heuristics and hence cognitive abilities that are in all likelihood not optimized to comprehend the essence behind the perceivable phenomena or the “absolute truth” in the Kantian sense. Evolution did not prepare us to directly perceive and understand processes at subatomic or cosmic scales, because they were and are completely irrelevant for our daily struggle for survival. Even more worrying is the possibility that the way in which we reason may also be limited by adaptation to those processes in the narrow range of the world that are relevant for survival and that we can access with our highly selective, specialized senses. In conclusion, it is very likely that our cognition is constrained. And this may apply not only to primary perception, but also to our way of deriving inferences from observables. If this were true it would pose unsurmountable barriers to our attempts to understand, just as it would challenge the consistency of mathematical theories and logical deductions. However, for these very reasons we have no way of knowing whether this is the case.

3.2 The contribution of neuroscience to epistemology

Growing insights into the neuronal mechanisms underlying perception provide compelling support for constructivist positions and emphasize the *epistemic caveats* formulated above. In the

light of neurobiological evidence, perceiving is essentially a constructive process. The sensory categories, for example those according to which we assign qualities to our experiences, are nothing but the idiosyncratic consequence of the layout of our sensory organs. These sample in a highly selective way a narrow range of physico-chemical signals, and this leads to the arbitrary classification of electromagnetic radiation with wavelengths between 400 to 700 nanometers as light, because the photoreceptors in the eye are sensitive to this wavelength range. Radiations with slightly longer wavelengths stimulate our temperature receptors and we categorise the respective sensations as temperature. A similar arbitrariness of category boundaries is observable in other sensory domains. The definition of perceptual objects, for example, is guided by a set of Gestaltrules that our brains apply in order to segment the spatio-temporal continuum of sensory signals into distinct objects—and this holds true for all sensory modalities. Objects are identified as such if they are delineated by spatial or temporal borders and exhibit some intrinsic coherence. This definition is appropriate in the mesoscopic world, but it does not apply to objects at atomic or subatomic scales. If we had no *a priori* definition of the properties of objects, we would not be able to distinguish objects, we would, for example, be unable to extract object-specific features from the continuous two-dimensional brightness distribution that cluttered scenes generate on the retina.

It is now well established by experimental evidence that the sparse sensory signals provided by our highly selective senses are interpreted by the brain on the basis of a vast amount of *a priori* knowledge that is stored in its own functional architecture. Our self-active brains permanently formulate knowledge and context-dependent expectancies, interpret sensory signals as a function of these inferences, and present the result of this constructive process to the workspace of consciousness. Paradoxically, we perceive the world around us as coherent even though our senses extract only a minute fraction of the available signals. Much of what we experience as actually perceived is read out from memory and is the result of reconstruction

and completion. This raises the question of the origins of this knowledge.

3.3 The sources of *a priori* knowledge

It is commonly accepted that all the knowledge a brain can possibly have, and the rules according to which this knowledge is applied for the interpretation of sensory signals and the execution of movements, reside in the functional architecture of the brain. This contradicts the analogy frequently drawn between computers and brains. Computers have processors and separate memories for programmes and for data. In the brain, however, there exist only neurons and connections. Both the stored knowledge and the programs for processing this knowledge reside in the layout of these connections, their polarity—that is, whether they are excitatory or inhibitory—and their graded efficacy. The question of the origin of stored information is thus reduced to the question of which processes determine the functional architecture of the brain.

The most important determinant of the functional architecture of brain—and hence the most important source of knowledge—is, of course, evolution. What makes our brain architectures comparable is evolutionary-acquired information that resides in the genes and determines the layout of the brain's connectome. It is knowledge about the world that is expressed in the functional architecture of brains every time an organism develops. In this sense evolution can be considered a cognitive process. This evolutionary-acquired knowledge pertains essentially to the conditions of the precultural world; and it is implicit—we are not aware of having it because we were not around when it was acquired. Still, we use it to interpret the signals provided by our sense organs and to structure adapted responses.

This inborn knowledge is subsequently complemented by extensive epigenetic shaping of the neuronal architectures, which adapt the developing brain to the actual conditions in which the individual develops. The human brain develops the majority of its connections only after birth, and this process continues approximately until the age of twenty or twenty-five

years. During this developmental period numerous new connections are formed, while many existing connections are removed; and this making and breaking is guided by the neuronal activity itself. Since, after birth, neuronal activity is modulated by interactions with the environment, the development of brain architectures is thus determined by a host of epigenetic factors derived from the natural and social environment. Through this process, the brain acquires knowledge about the specific conditions in which the newborn organism actually evolves, and thereby complements its genetically-inherited knowledge.

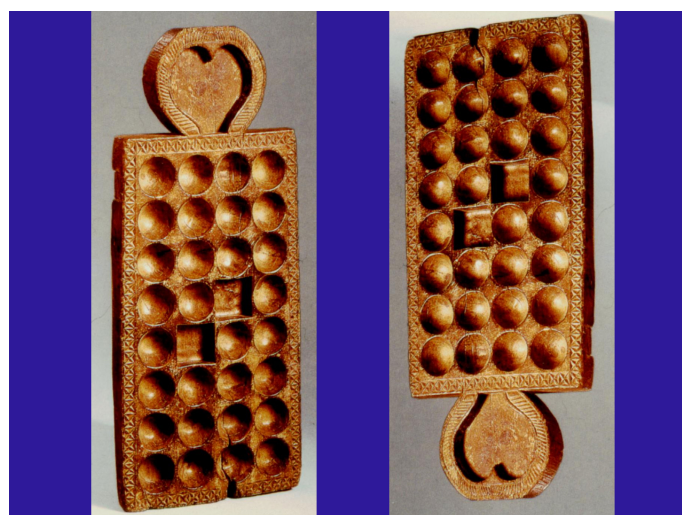


Figure 1: The brain assumes that light comes from above. The circular contours on the left board appear as concavities because the shadows are located at the right upper corner. The right board is actually the same as on the left, just rotated by 180 degrees. Now the shadows are on the lower left border and the contours appear convex.

A considerable part of this developmentally-acquired knowledge also remains implicit because of the phenomenon of childhood amnesia. Children before the age of about four years have only a limited capacity to remember in which context they have experienced and learnt particular contents. The reason for this is that the brain centres required for these storage functions—we call them episodic or biographical or declarative memories—have not yet matured. Thus, while young children learn very efficiently and store contents in a very robust way through structural modifications of their brain architec-

ture, they often have no recollection of the source of this knowledge. Because of this apparent lack of causation, the knowledge acquired in this way is implicit, similarly to evolutionary-acquired knowledge, and therefore often assumes the status of convictions that cannot be questioned.

Like innate knowledge, this acquired knowledge is used to shape cognitive processes and to structure our perceptions. Yet we are not aware that what we perceive is actually the result of such knowledge-based interpretations.

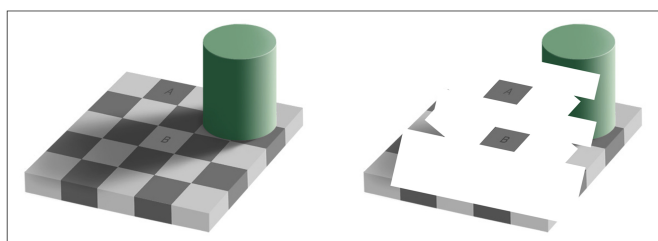


Figure 2: The checker-board illusion by Adelson, illustrating that even brightness perception depends on assumptions derived from context. (For further descriptions see text.)

Finally, there is knowledge acquisition by learning, which accompanies us throughout our lives. This is based on graded changes of the coupling strength of the existing connections between neurons. In the adult brain few new connections are formed and under normal conditions no breaking of connections occurs. The knowledge acquired by these learning processes also biases perception, but this is explicit, and its origins are known. One is usually aware of when and how it has been acquired and can therefore question its validity and by the same token the validity of what is perceived.

3.4 Examples illustrating the influence of priors on perception

The two examples depicted in [figure 1](#) and [figure 2](#) illustrate impressively how *a priori* knowledge structures our primary perceptions. The object in [figure 1](#) is a mould used to produce candies. On the left side one sees the inside of the mould, with its concavities, and on the left we see the rear side with its corresponding con-

vex protrusions. In reality, the pictures are identical, but one is rotated by 180°. The reason for our very different perceptions of the images is that the brain makes the *a priori* assumption that light comes from above—a well adapted assumption in a precultural world with only natural light sources. In this case contours that have the shadow above are interpreted as concave, and those with the shadow below as convex. Thus, an assumption of which we are not aware determines what we perceive. Another striking example is shown in [figure 2](#). It is hard to believe, but surfaces A and B have exactly the same luminance. They appear different because the brain sees the shadow that is caused by the cylinder on the right. Even though the amount of light reflected from surfaces A and B and impinging on the retina is exactly the same, the brain interprets the brightness of the two surfaces as different because it infers the following. Given that there is a shadow, surface B must be brighter than surface A—which has no shadow on it—in order to reflect the same amount of light. Thus, the brain “computes” the inferred brightness of the surfaces, but we are not aware of these computations. We just perceive the result and take it to be real, i.e., we see B as being much brighter than A. These two examples indicate that the brain generates inferences of which we are not aware, that it is permanently reconstructing the world according to *a priori* knowledge, and that we, as perceiving subjects, have to take for granted what the system finally offers us as conscious experience. As expected, this is not only the case with specially designed psycho-physical experiments, but is an essential feature of all our perceptual processes.

The mechanism leading to this “false” perception, to this “illusion”, has of course an important function. Our brain uses this principle to generate perceptual constancy, e.g., to keep colours and contrasts constant despite different illumination conditions. The spectral composition and the intensity of the sunlight change dramatically throughout the day, and therefore the spectral mix of light reflected from a particular, edible berry differs in the morning from that at noon. An animal that relies on colour to

distinguish one edible berry from another, slightly more violet and poisonous berry, cannot rely on an analysis of the “true” or actual spectral composition of reflected light. It first has to assess the spectral composition of the light source—the sunlight—and then must reconstruct the perceived colour. Our brains accomplish this by assessing the actual lighting conditions, by comparing the colours of the sky, of stones, of leaves and barks etc. and then, by using this contextual information, compute the “real” colour of the berries to identify that which is edible. Our brains are capable of assuring colour constancy despite changing illumination conditions, but we are completely unaware of the complexity of the computations assuring constancy and thereby survival in a changing world. In essence, all these operations are based on the evaluation of relations. We rarely perceive absolute values such as those measured by physical devices, be it intensities of stimuli, wave-lengths of sound or light waves, or chemical concentrations. We mostly perceive these variables in relation to others, as differences, increments and contrasts, such that these comparisons are made both across space and time. This is a very economical and efficient strategy because it emphasizes differences, permits coverage of wide ranges of intensities and, as mentioned above, allows for constancy. Given the advantages of these well-adapted heuristics it is at least questionable whether one should confront the resulting perceptions as illusions.

3.5 Conclusion of the excursion into epistemology

Evolution- and experience-dependent development determine and shape the architecture of the brain. Through these processes, knowledge about the world and strategies to use this knowledge for survival and reproduction are implemented in brain architectures. These in turn determine what and how an organism perceives and how it behaves. Because of the selection criteria that guides evolution, the brain adapted to the narrow segment of the world in which life has evolved, and its functions have been optimized to extract and process those signals that

best serve survival and reproduction. Thus the cognitive functions of the brain have probably not been optimized for understanding the deeper structure of the world that assures coherence across scales and cannot be perceived directly. Similarly, the rules according to which we evaluate contingencies and establish associations among events are implemented by specific molecular mechanisms that translate temporal correlations of neuronal activity into lasting changes in the efficacy of neuronal connections. These rules have been preserved virtually unchanged since the evolution of primitive nervous systems, and are at the basis of assignments of causality and the formation of associations. Again, these rules are highly efficient for the generation of models of the mesoscopic world and the formulation of predictions, but they do not apply to processes in the quantum world or to the relativistic dynamics of the universe. Given these specific adaptations of our cognitive functions, one might consider that similar restrictions may also hold for the way we reason. If so, this would present a serious challenge for the generalisation of models and theories based on extrapolation.

4 The contribution of the neurosciences to theories of consciousness

Some of the propositions summarized in the following chapter have been derived from experiments on the neuronal substrate of consciousness that have been described in detail in [Meloni & Singer \(2011\)](#) and [Aru et al. \(2012a, 2012b\)](#). A few decades ago, attempts to identify the neuronal correlates of consciousness (NCC) were considered futile. Because of the rapid development of non-invasive technology for the registration of neuronal activity in the human brain, and because of advances in the analysis of the neuronal underpinnings of higher cognitive functions, the search for NCC has now become a very active field of research in cognitive neuroscience. As expected, this new field is confronted with great challenges that are difficult to overcome. The explanandum is ill-defined; the prerequisites and consequences of conscious processing cannot easily be distinguished by ex-

periment from conscious processing per se; the experience of mental causation and agency is difficult to reconcile with contemporary concepts of self-organization; and finally it is difficult to bridge the epistemic gap between phenomena that are experienced from a first-person perspective and mechanisms described from a third-person perspective. Some of these problems will be addressed in the following paragraphs.

4.1 An attempt to define the explanandum

Most languages have coined a term for consciousness. Thus, it must be a robust phenomenon on which human beings can agree. However, while it is easy to use the term, it is virtually impossible to give a formal definition of what exactly it means. Nevertheless, the implicit understanding of what it is to be conscious seems to be sufficiently clear and widely accepted enough to justify a search for its neuronal correlates and, ultimately, to identify the neuronal mechanisms that enable a subject to be conscious of something. In their seminal paper, [Crick & Koch \(1990\)](#) propose that consciousness is a specific cognitive function and that, as such, it must have neuronal correlates that can be analyzed with the tools of the natural sciences. With the development of non-invasive imaging technologies, the tools became available to actually pursue this project and the search for the neuronal correlates of consciousness (NCC) became a mainstream endeavour.

Before discussing some of the proposed theories for NCC, I shall attempt to give an operational definition of what I mean when referring to awareness and consciousness or, in other terms, what it means to be aware of something or to be conscious. Subjects will be considered aware of something if they are able to report the presence or absence of the content of a cognitive process—irrespective of whether this content is made available by recall from stored memories or drawn from actual sensory experience. Thus, one criterion for awareness is the reportability of the presence of a cognitive content. These reports can in principle consist of any motor response, but to be on the safe side,

it is often requested that the report be verbal. The reason for this is that behavioural responses can be obtained under forced choice conditions that clearly indicate that the brain has processed and recognized the respective sensory material and produced a correct response even though the subject may not have been aware of having perceived the stimulus. There is thus an inherent ambiguity in non-verbal responses. They can but need not necessarily signal awareness, and this constrains research on NCC in animal experiments. Since consciousness is so difficult to define, an attempt will be made to avoid this term. Instead we shall use the adverb “consciously” and the adjective “conscious” in order to further specify particular brain states or aspects of a perceptual process. In addition, research into the hard problem of consciousness ([Chalmers 2000](#)) confronts the problem of explaining the phase transition from neuronal processes to the qualia of subjective experience, but this will be discussed only briefly at the end of this paper—and there in an enlarged context that transcends neurobiological approaches by also taking social interactions into account.

The state of being aware of something has a number of distinct properties that constrain the underlying neuronal mechanisms. One important feature of this state is unity or relatedness: Contents of which one is aware are experienced as simultaneously present and related to each other. Because of the distributed organization of brain processes, mechanisms supporting phenomenal awareness must therefore be able to bind together computational results obtained in multiple specialized and widely distributed processing areas. Another feature of awareness is that the contents that one is aware of change continuously but are bound together in time, appearing as a seamless flow that is coherent in space and time. Finally, subjects are only aware of a small fraction of on-going cognitive operations. Still, even signals of which subjects are not aware are often readily processed and impact behaviour ([Dehaene et al. 1998](#); [van Gaal et al. 2008](#)). Thus there must be gating mechanism that determines which signals are processed consciously, which are processed and control be-

haviour but remain unconscious, and which are excluded from processing altogether. Therefore, the identification of NCC requires a clear delineation between subconscious and conscious processes and an analysis of the mechanisms that gate access to awareness.

4.2 Conscious versus subconscious processing

As mentioned above, an enormous amount of knowledge is stored in the specific architectures of the brain, but we are not aware of most of these “given” heuristics, assumptions, and concepts. These routines determine the outcome of cognitive processes, which often have access to conscious recollection while themselves remaining hidden in the unconscious. We cannot move these implicit hypotheses and rules to the workspace of consciousness by focusing our attention on them, as is possible with most sensory signals and contents stored, for example, in “declarative memory”—the memory in which is inscribed what has been consciously experienced. Excluded from conscious experience are also certain sensory signals—such as those elicited, for example, by pheromones, which are processed by special olfactory subsystems—or the many signals from within the body—such as messages about blood pressures, sugar levels, and so on. It cannot be emphasized enough, however, that signals that are permanently excluded from conscious processing, as well as the facultatively-excluded signals from non-attended sensory stimuli, still have a strong impact on behaviour. Moreover, by influencing attentional mechanisms they can determine which of the stored memories or sensory signals will be transferred to the level of conscious processing. A hungry predator will search for traces of prey rather than mating partners, and so on.

One reason for the gated access of cognitive material to the level of awareness appears to be the limited capacity of the workspace of consciousness. Whether these limitations are due to the inability to attend to large numbers of items simultaneously, or whether they result from the restricted capacity of working memory, or even both, is subject to intense scientific investiga-

tion. The capacity of working memory is limited to about four to seven different items. The phenomenon of “change blindness”, which is the inability to detect local changes in two images presented in quick succession, demonstrates impressively our inability to attend to and consciously process all features of an image simultaneously. Because of these capacity constraints, conscious processing is in essence serial. Items are scrutinized and compared serially and therefore conscious processing is slow. Complex visual scenes are scanned serially and much of what we believe that we perceive simultaneously is actually reconstructed from memory. Which of the many signals finally reach the level of conscious awareness and can then be recalled depends on whether they are attended to, and this in turn is controlled either by external cues, such as the saliency of a stimulus, or by internal motifs, many of which we may actually not be aware of. And then it may occur that even an attentive, conscious search for content stored in declarative memory fails to raise it to the level of awareness. We are all familiar with the temporary inability to remember an episode or a name and have witnessed how a persisting subconscious search process suddenly lifts the content into the workspace of consciousness. It appears that we are not capable of controlling, at all times, which contents enter consciousness.

The differences between conscious and subconscious processes are further emphasized by evidence that the rules governing conscious deliberations and decisions most likely differ from those of subconscious processes. The former are based mainly on rational, logical, and syntactic rules, and the search for solutions is essentially based on serial computations. Arguments and facts are scrutinized one by one, and possible outcomes investigated. This strategy is suitable if variables are well defined, if sufficient time is available, if problems have a structure amenable to analytical treatment, and if precise solutions are required.

Subconscious mechanisms, by contrast, seem to rely more heavily on parallel processing, whereby a large number of variables enter into competition with one another. Then, a “winner takes all” algorithm leads to the sta-

bilization of the activity pattern that is the most likely, given the initial conditions and the heuristics derived either from inborn routines or from past experience. The domains of subconscious processing are situations requiring very fast responses or conditions where large numbers of underdetermined variables have to be considered simultaneously, and weighed against variables that have no or only limited access to conscious processing—such as the wealth of implicit knowledge and heuristics, vague feelings, hidden motives, or drives. The outcome of such subconscious processes manifests itself either in immediate behavioural responses or in what is called “gut feelings”. And it is often not possible to indicate with rational argument why exactly one has responded in such a way and why one feels that something is wrong or right. In experimental settings one can even demonstrate that the rational arguments given for or against a particular response do not correspond to the “real” causes. For the solution of complex problems with numerous entangled variables it often turns out that the subconscious processes lead to better solutions than conscious deliberations—and this is thought to be because of the wealth of heuristics exploitable by subconscious processing. Given the large amount of information and implicit knowledge to which consciousness has no or only sporadic access, and given the crucial importance of subconscious heuristics for decision-making and guidance of behaviour, first identifying the structure of a problem and then deciding whether one should rely on conscious deliberations or listen to the voices of the subconscious appears to be a well adapted strategy.

However, because the two systems operate according to different principles, the solutions to a particular problem may not always agree. Most of the decisions that get us through daily life rely on subconscious processing and follow well-adapted heuristics. If these decision processes do not lead to immediate action, they may still influence subsequent behaviour by manifesting themselves as what we call “gut feelings”. One has no conscious recollection of the reasons that lead to these feelings, but one clearly experiences the reactions of one’s

autonomous nervous system when the results of subconscious processes are in conflict with the outcome of conscious deliberations. In such situations one tends to say: “I decided according to all the rational arguments that I was aware of and took the best decision I could think of, but it somehow feels wrong.” The opposite situation is also possible, “I did what *felt* right to me, but if I think about it, it is absolutely crazy and irrational”. It is only when the two decision systems converge on the same solution one feels good, satisfied, and to some extent “free”.

After this brief excursion into the phenomenology of conscious and subconscious processes, intended to convey some connotations of consciousness, some of the most popular hypotheses about the constitution of consciousness in the brain will be reviewed.

4.3 Some competing hypotheses about the NCC

One class of theories focuses on the philosophical implications of the hard problem of consciousness without attempting to provide detailed descriptions of putative neuronal mechanisms (Searle 1997; Metzinger 2000; Dennett 1992; Chalmers 2000). Solutions to the hard problem have also been sought through transcending current concepts of neuronal processes and incorporating theories borrowed from other scientific disciplines. The most prominent of these approaches assumes that phenomena unravelled by quantum physics also play a role in neuronal processes, and that they might be able to account for the emergence of consciousness from material interactions in the brain (Hameroff 2006; Penrose 1994). None of the predictions of these theories are at present amenable to experimental verification, because there is no evidence that quantum phenomena such as entanglement, superposition and collapse of wave functions, etc., play a role at the macroscopic level of neuronal network functions. Quantum effects do of course exist at the level of molecular and submolecular interactions, but it appears highly unlikely that they are relevant for the macroscopic functions of neurons responsible for information processing. Thus quantum

theories of consciousness attempt to explain one poorly understood phenomenon with still unexplored and unproven mechanisms, and will therefore not be discussed further here.

Another class of theories pursues more modest goals and attempts to examine neuronal mechanisms potentially capable of supporting awareness of cognitive contents. Their aim is to define the neuronal mechanisms supporting the unitary character of awareness, its coherence in space and time, and the control of states that distinguish between conscious and unconscious processing (for review of relevant experimental findings see [Melloni & Singer 2011](#); [Aru et al. 2012a](#)).

The most intuitively plausible solution for the unity of awareness is convergence of the results obtained in distributed processing areas to a singular structure at the top of the processing hierarchy. Theories derived from this intuition predict the activation of specific cortical areas when subjects are aware of stimuli. Consequently, these regions should remain inactive during unconscious processing of the same material. Likewise, lesions of these putative areas should abolish the ability to become aware of perceptual objects. So far, a region with such universal “observer functions” has not been identified, and this option is considered theoretically implausible by some ([Dennett 1992](#)). There is also little—if any—experimental evidence for such a scenario. If brain lesions abolish the functions of sensory areas or regions involved in the recall of memories, patients lose the ability to consciously experience the respective sensory contents or memories, but the ability to process other material consciously remains unaffected. Moreover, behavioural and brain imaging studies have shown that unconscious processing engages very much the same areas as conscious processing, including the frontal and prefrontal cortices ([Lau & Passingham 2007](#); [van Gaal et al. 2008](#)). Thus there is no compelling evidence for specific areas supporting conscious processing. There are prominent examples of the selective elimination from conscious perception of those aspects of the stimulus material that are processed in specific regions without affecting awareness of other con-

tents, such as syndromes of agnosia and blindsight ([Cowey & Stoerig 1991](#)), which result from selective lesions of sensory subsystems.

There are, however, systems and pathways in the brain whose destruction abolishes all conscious experience—but these cannot be considered to be the NCC. Rather, these systems adjust the narrow dynamic range within which the brain has to be kept in order to be operational and to perform the computations that ultimately give rise to awareness. These systems are addressed as modulatory systems; they originate mainly in deep structures of the brain and control global brain states via widely-diverging ascending projections.

Another class of theories favours the notion that the mechanisms supporting awareness of stimulation material are distributed and do not require anatomical convergence ([Rodriguez et al. 1999](#); [Metzinger 2000](#); [Varela et al. 2001](#)). [Baars \(1997\)](#) and [Dehaene et al. \(2006\)](#) propose that there is a workspace of consciousness whose neuronal correlate is a widely distributed network of neurons located in the superficial layers of the cortical mantle. As mentioned above, these neurons are reciprocally coupled through a dense network of cortico-cortical connections that have features of small-world networks. The proposal is that subjects become aware of signals if these are sufficiently salient to ignite coordinated activity within this workspace of consciousness. This is assumed to be the case for signals that either have high saliency because of the high physical energy of the stimuli or those that are made salient due to attentional selection.

Yet another, related proposal is that subjects become aware of contents, irrespective of whether they are triggered by sensory events or recalled by imagery from stored memories, if the distributed neurons coding these contents are organized into assemblies characterized by coherent, temporally-structured activity patterns. In this case, the critical state variable distinguishing conscious from non-conscious processing would be the spatial extent and the precision of coherence of temporally-structured neuronal responses ([Rodriguez et al. 1999](#); [Metzinger 2000](#); [Varela et al. 2001](#)).

In what follows, evidence will be reviewed in support of the latter hypothesis. However, before discussing this evidence we should briefly recall the reasons why temporal coherence should matter in neuronal processing.

4.4 The formation of functional networks by temporal coordination

Because of the small-world architecture of the cortical connectome, any neuron can communicate with any other neuron either directly or via just a few interposed nodes. Thus, efficient and highly flexible mechanisms are required, which permit selective routing of signals and assure that only the neurons that need to interact in order to accomplish a particular task effectively communicate with one another. Evidence from multisite invasive recordings and from non-invasive registration of global activity patterns with magneto-encephalography or functional magnetic resonance imaging indeed indicates that functional sub-networks are configured “on the fly” on the backbone of fixed anatomical connections in a task- and goal-dependent way. One mechanism that can accomplish such fast and selective association of neurons and gate neuronal interaction is the temporal coordination of oscillatory activity (Gray et al. 1989; Fries 2005). Since the discovery (Gray & Singer 1989) that spatially-distributed neurons in the primary visual cortex tend to engage in oscillatory responses in the beta and gamma frequency band when activated by appropriately configured contours, and that these oscillatory responses can synchronize over large distances within and across cortical areas and even hemispheres, numerous studies have confirmed that oscillations and their synchronization in different frequency bands are an ubiquitous phenomenon in the mammalian brain. The pace-makers of these oscillations are reciprocal interactions in local networks of inhibitory and excitatory neurons. The long-distance synchronization of this oscillatory activity appears to be achieved by several mechanisms operating in parallel: Long-range excitatory cortico-cortical connections, long-range inhibitory projections, and pathways ascending from nuclei in the thal-

amus and the basal forebrain (for a review of these see Uhlhaas et al. 2009). When neurons engage in oscillatory activity, they pass through alternating cycles of high and low excitability. At the peak of an oscillation cycle neurons are depolarized, highly susceptible to excitatory input, and capable of emitting action potentials. In the subsequent trough of the cycle, the membrane potential is hyperpolarized and membrane conductance is high because of strong GABAergic inhibition generated by the rhythmically-active inhibitory interneurons. During this phase, neurons are less susceptible to excitatory inputs, because excitatory postsynaptic potentials (EPSPs) are shunted and because the membrane potential is far from the threshold. Hence, neurons are unlikely to respond to presynaptic excitatory drive.

These periodic modulations of excitability can be exploited in order to gate communication among neurons. By adjusting oscillation frequency and phases of coupled neuronal populations, communication among those neurons can either be facilitated or blocked. To form a functional network of distributed neurons it suffices to coordinate their oscillatory activity in such a way that signals emitted by neurons of this network impinge on other members of the network at times when these are highly susceptible to input. One way to achieve this is to entrain the neurons that should be bound into a functional network to engage in oscillations of the same frequency, to synchronize these oscillations, and to adjust the phases such that neurons that ought to be able to communicate can communicate.

Evidence from multi-site recordings indicate that neurons are indeed bound together into sometimes widespread functional networks through synchronization of their oscillatory activity in a task-dependent way (Salazar et al. 2012; Buschman et al. 2012). This supports the hypothesis (Gray et al. 1989; Singer 1999; Fries 2005) that synchronization of oscillatory neuronal activity is a versatile mechanism for the temporary association of distributed neurons and the binding of their responses into functionally coherent assemblies—which as a whole represent a particular cognitive content. Such a dy-

dynamic binding mechanism appears to be an economical and highly flexible strategy for coping with the representations of a virtually unlimited variety of feature constellations characterizing perceptual objects. Taking the unified nature of conscious experience and the virtually infinite diversity of possible contents that can be represented, the formation of distributed representations, through response synchronization, offers itself as a mechanism that allows for the encoding of ever-changing constellations of contents in a unifying format.

Synchronization is also ideally suited to contribute to the selection of contents for access to consciousness. It enhances the saliency of signals by concentrating spike discharges into a narrow temporal window. This increases the coincidence of excitatory postsynaptic potentials (EPSPs) in target cells that receive input from synchronized cell groups. Because coincident EPSPs summate much more effectively than temporally dispersed EPSPs, synchronized inputs are particularly effective in driving postsynaptic target cells. It is thus not unexpected that entrainment of neuronal populations in synchronized gamma oscillation is used for attention-dependent selection of input configurations (Fries et al. 2001b; Fries 2009).

4.5 A prediction relating long-range synchronisation to consciousness

If activation patterns that subjects can become aware of are indeed characterized by globally coherent states of those cortical regions that process the contents actually appearing as unified, we expect these states of awareness to be associated with large-scale synchronization of neuronal activity. Candidate frequency bands are gamma and beta oscillations, as these have been shown to serve the temporal coordination of cortical networks. By contrast, if subjects are not aware of the presented stimulus material, processing should remain confined to smaller sub-networks, which operate in relative isolation and are not integrated into globally coherent states. In this case one should observe only local synchronization of more circumscribed neuronal populations (see Varela et al. 2001).

Finally, adjustments of oscillation frequency and phases fulfil the requirement that assemblies representing consciously processed contents need to be reconfigured at an extremely fast rate. The contents of which subjects are aware can apparently change at a rapid pace, at least four times a second, if one considers that this is the frequency with which the direction of gaze changes during the scanning of natural scenes. Thus, assemblies representing contents that are consciously perceived must be reconfigurable at similarly fast timescales. Evidence suggests that cortical networks operate in a regime of self-organized criticality close to the edge of chaos (Shew et al. 2009). Dynamical systems operating in this range can undergo very rapid state changes, which are characterized by shifts in oscillation frequencies, synchronization, and phase.

4.6 Methodological caveats in the search for the NCC

As discussed previously (Aru et al. 2012a) experiments designed to identify the neuronal correlates of consciousness are often fraught with ambiguities. The most frequently used strategy for the identification of NCC is contrastive analysis. One creates perceptual conditions in which targets are consciously perceived in only a subset of trials, while making sure that physical conditions are kept as constant as possible. This strategy implies that detection tasks have been designed, which operate close to the perceptual threshold. This can be achieved by reducing the physical energy of the stimuli or by masking them. While subjects are engaged in such detection tasks, neuronal responses are measured and then trials are sorted depending on whether the subjects did or did not perceive the stimulus. By subtracting the average responses obtained in the two conditions from one another, those neuronal responses that occur only in the condition of successful detection can be isolated, and these are then commonly interpreted as the neuronal correlate of conscious perception. This seemingly simple approach is not without ambiguity. Thus, noise fluctuations in afferent pathways are likely to lead to signi-

ficant differences in the available sensory evidence, especially because experiments are performed at the perceptual threshold. Therefore, those aspects of neuronal responses that truly reflect NCC may be contaminated by signals resulting from noise fluctuations at processing stages preceding those actually mediating awareness. In addition, once subjects have become aware of stimuli, there are a number of subsequent processing steps that need not necessarily be linked to NCC. These comprise the covert verbalization of stimulus material, the engagement of working memory, the transfer of information into declarative memory, and perhaps also the preparation of covert motor responses. The distinction between these various confounding factors is difficult because all the processes are intimately related to each other. A detailed discussion of this problem is given in [Aru et al. \(2012a\)](#). One distinguishing feature could be the latency of the electrographic signatures of these various processing steps. Noise-dependent fluctuations in sensory evidence should be manifested early on; responses related to NCC proper should have some intermediate latency; and the consequences of having become aware of a stimulus should have the longest latencies. In order to use these latencies as distinguishing criteria, it is of course required that we estimate the precise latency at which the mechanisms leading to conscious perception are likely to be engaged. Assuming that the time required to prepare and execute simple motor responses is generally constant, the interval of interest can be constrained and has been proposed as somewhere between 180 and a few hundred milliseconds, depending on the sensory modality and the difficulty of the detection task. Attempts to use latency criteria for the elimination of confounds is of course restricted to electroencephalographic and magnetoencephalographic data, and cannot be applied to results obtained with functional magnetic resonance imaging because of the limited temporal resolution of this technique.

Another option for the reduction of confounds is to combine manipulations that influence the conscious perception of a stimulus through different mechanisms and to compare

the electrographic responses between conditions ([Aru et al. 2012b](#)). We applied this strategy in investigations of patients with subdurally implanted recording electrodes located over the visual cortex. In one set of trials the visibility of stimuli, in this case faces, was manipulated by changing the sensory evidence of the stimulus material. In another set of trials visibility of the same stimuli was influenced by allowing the subjects to familiarize themselves with some of the stimuli. This also facilitated detectability, but now because of an expectancy-driven top-down process. The reasoning behind this was that neuronal responses reflecting NCC proper should be the same irrespective of whether stimuli are consciously perceived because of enhanced sensory evidence or because of top-down facilitation. As electrographic signature of interest we analyzed the neuronal activity in the gamma band. In a previous study ([Fisch et al. 2009](#)) had shown that category-specific gamma band responses in the visual cortex correlate with conscious perception. Conscious recognition leads to a phasic enhancement of the gamma-band response, thus supporting the notion that conscious perception arises locally within sensory cortices—which is in line with previous conclusions ([Zeki 2001](#); [Malach 2007](#)). In our study we found that the performance and the reports of the subject were clearly modulated both by changing sensory evidence *and* by prior knowledge of the stimuli, as expected; but the gamma-band responses solely reflected the sensory evidence. This suggests that the differential activation of specific areas of the visual cortex, in our case mainly the fusiform face area, reflect processes that prepare access to conscious perception but are not its substrate proper.

Another frequently used paradigm in the search for NCC is interocular rivalry. If the two eyes are presented with stimuli that cannot be fused into one coherent percept, subjects perceive only one of the two stimuli at a time, and these percepts alternate. There are various ways to label the stimuli presented to the two eyes, to trace the responses related to their processing in the brain, and then to see which brain structures have to get involved in order to

support conscious perception. Again, these studies have led to inconclusive results. Some claim that suppression of signals corresponding to the non-perceived stimulus occurs only at very high levels of visual processing, such as, for example, the temporal cortex, which is the highest stage of the ventral processing stream. The conclusion of these studies is that activation of this particular cortical network is a necessary prerequisite for conscious processing (Logothetis et al. 1996; Silver & Logothetis 2004). Others, by contrast, found diverging activity patterns already existing at the level of the thalamus and the primary visual cortex (Haynes et al. 2005; Fries et al. 1997). Recent correlations between the dynamics characterizing binocular rivalry and anatomical features of the primary visual cortex and the commissures linking the primary visual cortices of the two hemispheres provide compelling evidence that the rivalry phenomenon is based on processes occurring within V1 (Genc et al. 2014). However, none of these studies allows us to unambiguously locate the processes that lead to conscious perception. They only contribute to the identification of the earliest levels of processing, in which changes are detectable that correlate with conscious perception.

Interesting and of potential relevance for interpretations given in the next subsection is the observation that the access of sensory signals to conscious processing does not seem to be gated by modulation of the neurons' discharge rate, but rather by changes of the synchronization of their activity—at least in the early stages of processing. What matters is the degree of synchronicity of oscillatory activity in the gamma frequency range. Signals conveyed by well-synchronized neuronal assemblies have access to conscious processing, while signals conveyed by similarly active but purely synchronized neurons fail to do so (Fries et al. 2001c). Of interest in this context is the observation that stimuli access conscious perception more easily if they are attended to, and that attention enhances synchronization of neuronal responses in the gamma frequency band in early visual areas (Fries et al. 2001b). Again, however, this local increase in synchrony is likely to simply enhance

the saliency of the neuronal responses, facilitating their propagation across the cortical networks, and cannot per se be considered a neuronal correlate of consciousness.

4.7 Evidence relating long-range synchronization and consciousness

The results described in what follows were obtained in a study where we presented words that could be perceived in some trials and not in others (by adjusting the luminance of masking stimuli), and simultaneously performed electroencephalographic (EEG) recordings (Melloni et al. 2007). Several measures were analyzed: Time-resolved power changes of local signals; the precision of phase synchronization across recording sites, and over a wide frequency range; and event-related potentials (ERPs). A brief burst of long-distance synchronization in the gamma-frequency range between occipital, parietal, and frontal sensors was the first event that distinguished seen from unseen words at about 180ms poststimulus. In contrast, local synchronization was similar between conditions. Interestingly, after this transient period of synchronization, several other measures differed between seen and unseen words: We observed an increase in amplitude of the P300 ERP for visible words, which most likely corresponds to the transfer of information to working memory. In addition, during the interval period—in which visible words had to be maintained in memory—we observed increases in frontal theta oscillations. Theta oscillations have been related to maintenance of items in short-term memory (Jensen & Tesche 2002; Schack et al. 2005).

To test whether the increase in long-distance synchronization relates to awareness or depth of processing, we further manipulated the depth of processing of invisible words. It has previously been shown that invisible words can be processed up to the semantic and motor level (Dehaene et al. 1998). In a subliminal semantic priming experiment we briefly presented words (which were thus invisible) that were either semantically related or unrelated, alongside a second, visible word on which subjects had to carry out a semantic classification task. Invis-

ible words were processed up to semantic levels, as revealed by modulation of the reaction times, depending on the congruency between invisible and visible words: congruent pairs exhibited shorter reaction times than incongruent ones. We observed increases in power in the gamma frequency range for unseen but processed words. For visible words we additionally observed increases in long-distance synchronization in the gamma-frequency range (Melloni & Rodriguez 2007). Thus, local processing of stimuli is reflected in increases in gamma power, whereas long-distance synchronization seems to be related to awareness of stimuli. This suggests that conscious processing requires a particular dynamical state of the cortical network. The large-scale synchronization that we observed in our study could reflect the transfer of contents into awareness and/or their maintenance. We favour the first possibility, given the transient nature of the effect, and argue that the subsequent theta oscillations might support maintenance. It is conceivable that short periods of long-distance synchronization in the gamma band reflect the update of new contents, while the slower pace of theta oscillations might relate to sustained integration and maintenance of local results in the workspace of consciousness. The interplay between these two frequency bands might underlie the phenomenon of continuous but ever-changing conscious experience (see below).

More recently, Gaillard et al. (2009) have revisited the processing of visible and invisible words. In intracranial recordings in epileptic patients, they observed that invisible words elicited activity in multiple cortical areas, which quickly vanished after 300 ms. In contrast, visible words elicited sustained voltage changes, increases in power in the gamma band, as well as long-distance synchronization in the beta band and long-range Granger causality. In contrast to our study, Gaillard et al. observed a rather late (300–500 ms) rise of long-distance synchronization. However, it is important to note that in the study undertaken by Gaillard et al., phase-synchrony was analyzed for the most part over electrodes within a given cortical area, or at most between hemispheres. It is thus conceiv-

able that earlier synchronization events passed undetected because of incomplete electrode coverage. Despite these restrictions, this study provides one of the most compelling pieces of evidence for a relation between long-distance synchronization and consciousness.

Some results of the experiments on binocular rivalry point in the same direction. Several studies have shown increased synchronization and phase locking of oscillatory responses to the stimulus that was consciously perceived, and controlled behaviour (Cosmelli et al. 2004; Doesburg et al. 2005; Fries et al. 1997; Srinivasan et al. 1999). Cosmelli et al. (2004) extend the findings obtained in human subjects by performing source-reconstruction and analyzing phase-synchrony in source space. These authors observed that perceptual dominance was accompanied by co-activation of occipital and frontal regions, including the anterior cingulate and medial frontal areas. Recently, Doesburg et al. (2009) have provided evidence for a relation between perceptual switches in binocular rivalry and theta- and gamma-band synchronization. Perceptual switches were related to increments in long-distance synchronization in the gamma band between several cortical areas (frontal and parietal) that repeated at the rate of theta oscillations. The authors suggest that transient gamma-band synchronization supports discrete moments of perceptual experience, while theta oscillations structure their succession in time, pacing the formation and dissolution of distributed neuronal assemblies. Thus, long-range gamma synchronization locked to on-going theta oscillations could serve to structure the flow of conscious experience, allowing for changes in content every few hundred milliseconds. Further research is required to clarify the exact relation between the two frequency bands, their respective role in the generation of percepts, and the pacing of changes in perception.

Another paradigm in consciousness research exploits the attentional blink phenomenon. When two stimuli are presented at short intervals among a set of distractors, subjects usually detect the first (S1) but miss the second (S2) when it is presented 200–500 ms after S1. Increases in long-range neuronal synchrony in the beta and gamma frequency ranges have been observed when the S2

is successfully detected (Gross et al. 2004; Nakatani et al. 2005). Furthermore, Gross et al. (2004) observed that successful detection of both S1 and S2 was related to increased long-distance synchronization in the beta range to both stimuli, and this enhanced synchrony was accompanied by higher de-synchronization in the inter-stimulus interval. Thus, de-synchronization might have facilitated the segregation of the two targets, allowing for identification of the second stimulus (also see Rodriguez et al. 1999). Source analysis revealed, as in the case of binocular rivalry, dynamical coordination between frontal, parietal, and temporal regions for detected targets (Gross et al. 2004).

In summary, studies of masking, binocular rivalry, and the attentional blink support the involvement of long-range synchronization in conscious perception. Recent investigations have suggested further that a nesting of different frequencies, in particular of theta and gamma oscillations, could play a role in pacing the flow of consciousness. Furthermore, the study of Gross et al. (2004) suggests that de-synchronization could serve to segregate representations when stimuli follow at short intervals. These results are encouraging and should motivate further search for relations between oscillatory activity in different frequency bands and consciousness, whereby attention should be focused not only on the formation of dynamically-configured networks but also on their dissolution.

4.8 Conclusions on putative mechanisms supporting consciousness

Of the numerous proposals on NCC, those favouring temporal coherence as the mechanism that integrates widely distributed processes appear to be the least controversial. They account best for the apparent discrepancy between the unity of conscious experience and the distributed organization of the brain, because they allow for the dynamic integration of information generated in parallel by spatially segregated processing modules. Large-scale synchronization of oscillatory activity has been identified as a candidate mechanism for the flexible coupling of widely-distributed neuron populations, and hence as a likely NCC. This variable has the advantage that it can be meas-

ured relatively directly in humans who are able to give detailed descriptions about their conscious experience. However, oscillations and synchrony seem to be mechanisms that are as intimately and inseparably related to neuronal processing in general as the modulation of neuronal discharge rates. Thus, without further specification these phenomena cannot stand up as NCC—at least when we disregard the triviality that consciousness does not exist without them. We propose that the spatial scale and the precision and stability of neuronal synchrony might be taken as more specific indicators of whether the communication of information in the brain is accompanied by conscious experience or not. In this framework, conscious experience arises only if information that is widely distributed within or across subsystems is not only processed and passed on to executive structures but also bound together into a coherent, all-encompassing, non-local but distributed meta-representation. This interpretation is compatible with views that take consciousness to be the result of the dynamic interplay of brain subsystems; one that allows for a rapid and highly-flexible integration of information provided by the numerous distributed subsystems that operate in parallel. This view resembles a proposal from Sherrington, formulated in his book *The Integrative Action of the Nervous System* (Sherrington 1906): “[p]ure conjunction in time without necessarily cerebral conjunction in space lies at the root of the solution of the problem of the unity of mind.” The additive value of conscious processing would then be the possibility of establishing in a unified data format ever-changing relations between cognitive contents—irrespective of whether they are read out from memory or induced by sensory signals, and irrespective of the sensory modality providing the signals. By virtue of this dynamic definition of novel relations, non-local meta-representations of specific constellations could be established that have the status of cognitive objects. Just as with any other distributed representation of contents, these could then be stored as distributed engrams by use-dependent modification of synaptic connections, and thus influence future behaviour. Thus, conscious processing would differ from non-conscious processing because it allows for the versatile binding

of the previously unbound into higher-order representations. And if so, “conscious” processing would be functionally relevant and not merely an epiphenomenon. Further arguments supporting the functional role of conscious processing are presented in the following section.

5 Arguments supporting an adaptive value of conscious processing

5.1 Evolutionary considerations

Given the continuity of evolution and the gradual increase in complexity of brains that reached a (perhaps preliminary) maximum in human beings, it appears likely that the ability to be aware of cognitive contents, of one’s own cognitive operations, and finally of oneself as an intentional agent, is not an all-or-none phenomenon but an ability that gradually emerged as we evolved. Paleoanthropological evidence supports this hypothesis by documenting correlations between increasing brain volume and increasingly refined artefacts that reflect gradual increases in cognitive abilities. Not much direct data are available on the evolution of the human brain, because our immediate ancestors are all extinct. Thus, we have to rely on evidence of comparative studies with brains of other species. In less evolved brains the paths from sensory to executive areas of the cerebral cortex are short. These relatively short sensory-motor loops are of course much more elaborate than simple reflex loops, because signals are processed extensively and transmission is made conditional on input from other systems, on past experience, and on context. As evolution proceeds and brains become more and more complex, one observes the addition of new cortical areas. A distinctive feature of this evolutionary process is the way in which these new areas are embedded in already-existing networks. These newly-added areas no longer communicate directly with the periphery, neither on the executive nor on the receptive side (see [figure 3](#)). Instead they receive their input exclusively from the phylogenetically older areas, and also distribute their computational results solely to other cortical areas and not to effector systems. This process is iterative, with more and more areas constantly

added that communicate only with other higher areas. A neuron located in these more recent areas is connected exclusively with other partners in the cerebral cortex. Evidence indicates that all cortical areas, including older and more recent ones, operate according to the same basic principles, because they share the same intrinsic organization. These purely anatomical considerations suggest that the evolutionary-recent areas process the results of the older areas similarly to how they process signals from the outer world. This iteration of “cognitive” processes across several hierarchical levels could thus generate representations of representations, i.e., meta-representations. Information that has already been processed by the already existing areas becomes the object of yet another cortical computation, i.e., of a secondary cognitive operation. These iterative operations can even be circular, because all these areas are interconnected reciprocally. Thus, higher-order areas feed back to lower, order areas and can have their results reprocessed. In principle this recursive process should permit the generation of meta-representations of increasingly higher order. In other words, highly evolved brains can apply their cognitive functions not only to the outer world but also to processes that occur within the brain. Brain processes can therefore become the object of the brain’s own cognitive operations. This could be the basis of phenomenal awareness, the awareness of perceiving, to create a protocol for what one perceives, and in the case of human beings to create symbols for the perceived and for internal states and to communicate them to others. Animals probably share some of these abilities, because their brains are organized in a very similar way.

Curiously, the ability to be aware of the results of cognitive operations provides no clue concerning the computational processes underlying these cognitive functions. We have no insight into the neuronal processes that bring about cognition. We are aware only of the results—just as we are aware of an action without being able to tell which neuronal processes in the motor centers of our brains caused this action. This fact is at the origin of most discrepancies between our intuitions and neurobiological evidence concerning the nature of agency, the experience of Free Will, and

the ontological status of qualia and consciousness (see [final](#) paragraph).

These evolutionary considerations may provide some plausible explanation of the emergence of higher cognitive functions—including consciousness—but they do not suffice to counter the argument that the emergence of consciousness is an epiphenomenon without adaptive value. To address this problem, one must identify functions that can only be realized by conscious processing.

Emergence of phenomenal awareness

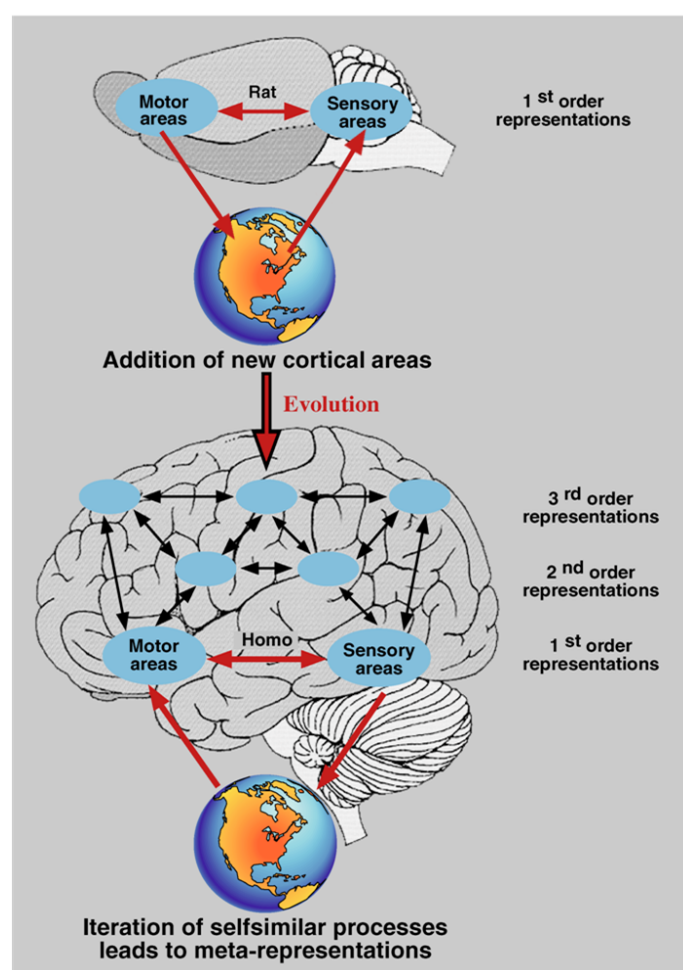


Figure 3: The evolution of complex brains is characterized by a massive increase of cortical areas. This renders responses to stimuli increasingly dependent on intracerebral processes and permits generation of meta-representations.

5.2 Functional considerations

Is the ability to be conscious of one's own cognitive operations a fitness factor? And would it

make any difference if brains lacked this ability? The philosopher of mind, David Chalmers, once stated: “[n]o, that wouldn’t make a difference. It’s just an epiphenomenon and if we wouldn’t have it we would do as well because the underlying brain processes would be the same and get us through life without us having to be aware of them.” I tend to disagree with this view for a number of reasons. The common denominator, however, is that there is something very special about the nature of conscious processes, and that this uniqueness does indeed constitute a fitness factor—in particular with respect to the ability to develop symbolic communication systems, a theory of mind, differentiated social systems, and, ultimately, culture.

As has been argued above, conscious processing allows for abstraction and symbolic representations due to its versatile binding of virtually all results of lower order cognitive operations in a unified representational space, capitalizing on the lingua franca (the homogeneous data format) of communication among cortical areas. This permits implementation of very effective—we call them rational—strategies for deliberation and decision-making that differ from and complement those of subconscious processes. The question then remains whether the results of these special processes affect brain functions and thereby affect behaviour.

It is difficult to see how the outcome of a conscious deliberation, that is, of an argument-based decision, could not affect future behaviour. A conscious decision leaves traces in declarative memory and so must the consequences of this decision. These traces are inscribed as modifications of the functional architecture of the respective networks. If a decision has averse or beneficial consequences, the experience of these consequences will also alter network properties, and the novel activation patterns generated by the modified networks will enter as a novel argument, or as a change of goal in future deliberations. Eventually, the newly-set goal, which initially had the status of a conscious rational argument, may change its status and become a habit that henceforth influences behaviour without having to appear as an explicit argument in consciousness. It can become one of

the variables that act at a subconscious level. One then refuses another glass of wine not because one recapitulates all the rational arguments against alcohol consumption but simply because it does not feel right to drink too much.

Another, and probably the most important fitness factor of conscious processing is its capacity to support complex societies in cultural evolution. This suggests that there may have been a co-evolution of mechanisms supporting the emergence of conscious behaviour on the one hand and the formation of societies on the other, with the two developments mutually supporting each other.

Cognitive objects represented in consciousness are always bound together into a coherent experience. Whether this is also the case for subconsciously-processed contents is obviously difficult to ascertain, but the following argument suggests that subconscious processing may be less integrated, more modular, and confined to subsystems. We subconsciously orient towards salient stimuli irrespective of their modality, visual, auditory, or tactile qualities, and these stimuli may be analyzed in the subconscious with respect to their behavioural relevance and thus give rise to action. However, if several stimuli compete for processing, usually the most salient will win. Compared to orienting responses, which are guided by conscious processing, there seems to be little evaluation of the embedding context here. Stimuli are processed more independently than they would have been had they entered consciousness, been bound together, and formed a unified coherent percept.

In order to achieve this integration and to relate the signals from various sensory systems to one another at the semantic level, the signals have to be encoded in a sufficiently abstract and homogenous format. As mentioned above, when describing brain evolution, the substrate for this integration of signals preprocessed by segregated sensory systems may be the evolutionary-recent cortical areas. By virtue of integrating and comparing signals from different modalities, it becomes possible to detect the similar in the seemingly different, and hence to extract invariant properties and to arrive at ab-

stract descriptions. Thus the addition of the novel, so called “association areas” of the neo-cortex prepared the ground not only for a more unified, polymodal representation of cognitive contents but also for symbolic coding—which in turn is a prerequisite for the development of a symbolic language system and abstract reasoning. Thus one might consider consciousness, or the state of conscious processing, as a state where distributed computational results can be bound together into a coherent whole, establishing multiple, simultaneous relations between the various distributed items. This obviously allows for a more abstract, more symbolic, and more comprehensive description of conditions. By itself this is an advanced processing strategy, whose adaptive or fitness value is obvious. However, if this unified, condensed, and abstract information can be routed to a versatile communication system, as seems to be the case with contents that are processed consciously, the evolution of cooperating societies will be greatly facilitated. Not only will it be easier to communicate what one has perceived if the numerous signals from different sensory modalities have already been bound together into coherent wholes but, because of the reflexive nature of awareness, it will also be easier to convey information about one’s internal state. This, in turn, is an important prerequisite for society-building, because it nurtures trust in and predictability of the respective other. A condensed report of the actual contents of one’s conscious state and its storage in the episodic memory of the listener is an effective and parsimonious way to couple brains with one another, to share experiences, and to foster cooperation. This interaction will modify the brains of the communicating partners and thereby act on their behaviour. In a sense this is an example of mental or top-down causation. The results of an information-processing strategy that can only be realized in the workspace of consciousness are stored in declarative memory—in this case not only in a single brain but in those of communicating partners—and henceforth influence future cognitive processes and behaviour. These considerations suggest that it might be useful and perhaps even necessary in consciousness re-

search to consider the phenomena ascribed to consciousness not solely in the context of cognitive functions of individual brains but in the larger context of social interactions. In the following section this strategy is applied in an attempt to approach the “hard problem of consciousness”.

6 Consciousness as a social phenomenon

6.1 Is the hard problem resolvable by considering cultural rather than only biological evolution?

The hard problem in consciousness research, the epistemic difficulty of devising bridging theories between subjective phenomena available only from a first-person perspective and the underlying neuronal mechanisms analyzable only from the third-person perspective, may not be resolvable by considering only the cognitive abilities of isolated brains. In addition, more recent concepts of embodiment that consider the embedding of the nervous system in a body endowed with receptor and effector organs may not suffice. Rather, it may be necessary to reconsider the problem in the context of social phenomena or social realities that emerged during cultural evolution. Through social interactions, realities have been created that can readily be experienced as such but that transcend the reality that existed before humans added cultural to biological evolution. These new realities have the quality of relations. They are immaterial, not tangible, not visible, not directly accessible to our senses—and yet they are perceived as real, as mental objects that humans can agree upon, that can become the object of shared attention and influence behaviour—just like the equally immaterial contents of belief systems.

What seems to pose the epistemic—the hard—problem in philosophy of mind is not so much that we perceive and have feelings and emotions, since we readily grant such abilities to animals and seem to be quite successful in identifying the neuronal substrate of these functions. The real problem appears to be our meta-awareness of having these abilities. It is

this meta-awareness that has mental connotations that appear to be so difficult to relate to neuronal processes. However, this meta-awareness—which is so intimately related with what we address when we talk about consciousness—in all likelihood does not emerge naturally from the functions of an individual brain. Rather, it appears to be the consequence of experiences resulting from social interactions, just as is the case for the experience of individuality, agency, and intentionality. These are attributions that have their roots in interpersonal interactions and are probably appropriated by individuals while they are developing their self-image. Without being embedded in a differentiated socio-cultural environment, without the option of mirroring oneself in the perception of others, without reflexive interactions between persons endowed with a theory of mind, and without an exchange of reports about inner states, formulated in a symbolic language-system, we would probably not be aware of our being conscious.

Thus the phenomenon that we call consciousness has the ontological status of a social reality. It is a construct, just like all the other social realities that our cultures have brought into this world. However, this construct differs in an important respect from other social realities, such as norms, beliefs, and values, in that it is an attribution that we ascribe to ourselves. We learn from social interactions that we are endowed with the ability to be aware of being aware. After conceptualization of this novel experience it becomes an integral part of our self-image—we exchange reports on this shared experience and coin words for it.

Regarded in this way the “hard problem” may not be as hard as it seems. Analyzing individual brains will not unravel the “correlates” of the many semantic connotations of what we term consciousness, because they emerged from a reflexive interaction *between individuals*. But we should eventually be able to identify the neuronal mechanisms underlying the cognitive abilities that allowed human beings to create the socio-cultural environment that itself allowed for the experience of being conscious.

7 Culture-specific, epigenetic shaping of brains and the concept of tolerance

As suggested in the preceding section, inclusion of the social dimension may be indispensable for defining the status of higher cognitive functions such as consciousness. But it may also help us to bridge gaps between naturalistic approaches and first-person phenomenology. Thus, surprisingly, the joint consideration of neurobiological evidence on the dependence of perception of priors and of the epigenetic modification of priors by socio-cultural factors has normative consequences for concepts of tolerance.

Human beings have a similar genetic outfit, and therefore usually agree on what they perceive—in particular when priors suffice for the interpretation of the perceived, which had been acquired during evolution in a pre-cultural world. However, this may not apply to the perception of social realities. Humans raised in different cultures may perceive social realities quite differently. Cultures may differ radically with respect to social conditions, concepts of fairness, justice, and aesthetics, as well as moral criteria, and so on. Thus, epigenetically-installed priors are likely to exhibit considerable culture-specific differences. As a consequence, the ways in which the world is perceived will also differ. Another culture specific variable likely to influence perception is the attachment to caretakers. The nature of early bonding experiences determines whether the world appears as hostile or as peaceful and secure, and whether others can be approached with confidence or scepticism. The ways in which these early attachments to other members are secured differ dramatically among different cultures and, as such, so will the perception of signals that inspire confidence or aggression. Since these early-acquired priors are implicit, and since one is unaware that perceptions are influenced by these early imprinting processes, human beings take what they perceive as “real” and see no reason to question its validity. Thus, two persons raised differently, while observing the same social situation may perceive it in completely different ways. They may come to grossly diverging ethical or moral judgments,

unable to convince the other through argument that he or she is wrong, because both experience what they experience as reality—just as one experiences optical illusions as real. The problem is that, in the case of the perception of social realities, there are no “objective” measuring devices. There are different perceptions, and there is no right or wrong. This has far-reaching consequences for our concepts of tolerance. Solving such problems with majority votes is clearly not a fair solution. Assuming that one’s own position is correct and granting others the right to retain their “wrong” perceptions—so long as they do not disturb the peace—is humiliating and denies them respect. Still, this is often considered to be tolerant behaviour. What should be done, instead, is to grant everybody that her or his perceptions are correct and to postulate that this attitude be reciprocated. Only if this agreement on reciprocity is violated have the dissenting parties the right to exert sanctions.

8 Concluding remarks

In this paper only a selection of concepts currently under discussion in the field of cognitive neuroscience could be reviewed in any detail. The selection criterion was their relatedness to consciousness studies and to some extent their relevance for epistemic issues. This led to the neglect of important domains, some of which have equally relevant bearing on philosophical questions and societal issues. Thus, no consideration was given to developmental aspects that illustrate the close correlation between the maturation of particular brain structures and the emergence of specific cognitive and social functions. Equally neglected were research on executive functions and associated concepts on intentionality, agency, mental causation, and free will. No attempts were made to explore the vast field devoted to the analysis of genuinely “psychic” phenomena, such as memories, emotions, motivations, drives, and aesthetic judgement. Finally, a much more elaborate review would have been required of research that analyzes brain functions that can only be assessed in conditions where brains interact with one an-

other in a social context. This is the domain of the relatively recent field of social neuroscience, which studies phenomena such as confidence, greed, generosity, fairness, parasitism, altruism, compassion, and collective beliefs. This would have been particularly important in view of the attempt made at the end of this contribution to demystify some of the immaterial connotations of consciousness, in particular by redefining some aspects of consciousness as belonging to the domain of social realities, i.e., as a relational construct emerging from reciprocal interactions among brains endowed with the cognitive abilities of humans. The research agendas of these “neglected” fields are the same as those of the domains reviewed above. They all attempt to identify the neuronal mechanisms that are responsible for a particular cognitive or executive function, such that it becomes less and less important whether a cognitive phenomenon is accessible from the third-person perspective—such as the orienting behaviour of an animal—or whether it is accessible only from the first-person perspective—such as in the famous example of the hue of a rose or an emotion. As long as these qualia can be operationalized and rated on some subjective scale, they are amenable to neuroscientific inquiry. In case of the social neurosciences, the agenda is somewhat enlarged, since the objects of studies are phenomena that emerge from social interactions and exist as relational constructs only in interpersonal space. Here the objects of study are the mechanisms underlying the cognitive functions enabling the respective social interactions such as, for example, the ability to have a theory of mind and the mechanisms that permit individual brains to represent social realities. This agenda of the neurosciences may appear bold and, as the reader will have noticed, while we already know much about the component functions in our brains, we are still very far from understanding the distributive processes underlying higher cognitive and executive functions. In fact, the more data that sophisticated analytical tools allow us to accumulate, the more we are humbled by the mind-boggling and no longer intuitively graspable complexity of the brain’s dynamics. However, at least the author and probably the ma-

jority of his colleagues believe that there should be no principle epistemological barriers to the pursuit of this research agenda. The greatest problem in the near future will be that the description of the dynamics of neuronal processes underlying higher functions will take the form of abstract mathematical formulations that lack any resemblance to the experienced or observed result of these functions.

Nevertheless, the naturalistic stance taken by the neurosciences has already in these early days provided some insights, whose relevance goes beyond the research agenda of the neurosciences proper. The data on mechanisms mediating perception discussed at the beginning of this chapter clearly support constructivism and thereby provide arguments for or against particular philosophical positions. Likewise, these epistemic considerations, to the author’s surprise, led to normative consequences in the context of notions of tolerance. Similar normative consequences arise from data on mechanisms responsible for decision-making, motivation, response suppression, conscious versus subconscious processing, personality traits, and so on, as these insights are all consequential for the definition of behavioural norms and the distinction between normal and pathological behaviour. It is foreseeable, therefore, that the neurosciences will become more involved in philosophical, normative, ethical, and societal issues. This should be beneficial for all parties involved, since the communication process is likely to lead to bridging theories, new terms for mutual understanding, and amendments to discipline-specific idiosyncrasies.

References

- Aru, J., Bachmann, T., Singer, W. & Melloni, L. (2012a). Distilling the neural correlates of consciousness. *Neuroscience and Biobehavioral Reviews*, 36 (2), 737-746. [10.1016/j.neubiorev.2011.12.003](https://doi.org/10.1016/j.neubiorev.2011.12.003)
- Aru, J., Axmacher, N., Do Lam, A. T., Fell, J., Elger, C. E., Singer, W. & Melloni, L. (2012b). Local category-specific gamma band responses in the visual cortex do not reflect conscious perception. *The Journal of Neuroscience*, 32 (43), 14909-14914. [10.1523/JNEUROSCI.2051-12.2012](https://doi.org/10.1523/JNEUROSCI.2051-12.2012)
- Baars, B. J. (1997). In the theatre of consciousness. Global workspace theory, a rigorous scientific theory of consciousness. *Journal of Consciousness Studies*, 4 (4), 292-309.
- Buffalo, E. A., Fries, P., Landman, R., Buschman, T. J. & Desimone, R. (2011). Laminar differences in gamma and alpha coherence in the ventral stream. *Proceedings of the National Academy of Sciences of the USA*, 108 (27), 11262-11267. [10.1073/pnas.1011284108](https://doi.org/10.1073/pnas.1011284108)
- Buonomano, D. V. & Maass, W. (2009). State-dependent computations: spatiotemporal processing in cortical networks. *Nature Reviews Neuroscience*, 10, 113-125. [10.1038/nrn2558](https://doi.org/10.1038/nrn2558)
- Buschman, T. J., Denovellis, E. L., Diogo, C., Bullock, D. & Miller, E. K. (2012). Synchronous oscillatory neuronal ensembles for rules in the prefrontal cortex. *Neuron*, 76 (4), 838-846. [10.1016/j.neuron.2012.09.029](https://doi.org/10.1016/j.neuron.2012.09.029)
- Buzsáki, G. (2006). *Rhythms of the brain*. Oxford, UK: Oxford University Press.
- Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C. R., McGuire, P. K., Woodruff, P. W. R., Iversen, S. D. & David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science*, 276 (5312), 593-596. [10.1126/science.276.5312.593](https://doi.org/10.1126/science.276.5312.593)
- Chalmers, D. J. (2000). What is a neural correlate of consciousness? In T. Metzinger (Ed.) *Neural correlates of consciousness: Empirical and conceptual questions* (pp. 17-40). Cambridge, MA: MIT Press.
- Constantinople, C. M. & Bruno, R. M. (2013). Deep cortical layers are activated directly by thalamus. *Science*, 340 (6140), 1591-1594. [10.1126/science.1236425](https://doi.org/10.1126/science.1236425)
- Cosmelli, D., David, O., Lachaux, J.-P., Martinerie, J., Garnero, L., Renault, B. & Varela, F. (2004). Waves of consciousness: ongoing cortical patterns during binocular rivalry. *NeuroImage*, 23 (1), 128-140. [10.1016/j.neuroimage.2004.05.008](https://doi.org/10.1016/j.neuroimage.2004.05.008)
- Cowey, A. & Stoerig, P. (1991). The neurobiology of blindsight. *Trends in Neurosciences*, 14 (4), 140-145. [10.1016/0166-2236\(91\)90085-9](https://doi.org/10.1016/0166-2236(91)90085-9)
- Crick, F. & Koch, C. (1990). Towards a neurobiological theory of consciousness. *Seminars in Neuroscience*, 2, 263-275.
- David, S. V., Vinje, W. E. & Gallant, J. L. (2004). Natural stimulus statistics alter the receptive field structure of V1 neurons. *The Journal of Neuroscience*, 24 (31), 6991-7006. [10.1523/JNEUROSCI.1422-04.2004](https://doi.org/10.1523/JNEUROSCI.1422-04.2004)
- Dehaene, S., Naccache, L., Le Clec', H. G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., van de Moortele, P.-F. & Le Bihan, D. (1998). Imaging unconscious semantic priming. *Nature*, 395 (6702), 597-600. [10.1038/26967](https://doi.org/10.1038/26967)
- Dehaene, S., Changeux, J. P., Naccache, L., Sackur, J. & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: A testable taxonomy. *Trends in Cognitive Sciences*, 10 (5), 204-211. [10.1016/j.tics.2006.03.007](https://doi.org/10.1016/j.tics.2006.03.007)
- Dennett, D. C. (1992). *Consciousness explained*. London, UK: Penguin.
- De Pasquale, R. & Sherman, S. M. (2011). Synaptic properties of corticocortical connections between the primary and secondary visual cortical areas in the mouse. *The Journal of Neuroscience*, 31 (46), 16494-16506. [10.1523/JNEUROSCI.3664-11.2011](https://doi.org/10.1523/JNEUROSCI.3664-11.2011)
- Doesburg, S. M., Kitajo, K. & Ward, L. M. (2005). Increased gamma-band synchrony precedes switching of conscious perceptual objects in binocular rivalry. *Neuroreport*, 16 (11), 1139-1142. [10.1097/00001756-200508010-00001](https://doi.org/10.1097/00001756-200508010-00001)
- Doesburg, S. M., Green, J. J., McDonald, J. J. & Ward, L. M. (2009). Rhythms of consciousness: Binocular rivalry reveals large-scale oscillatory network dynamics mediating visual perception. *PLoS One*, 4 (7), e6142. [10.1371/journal.pone.0006142](https://doi.org/10.1371/journal.pone.0006142)
- Engel, A. K., Fries, P. & Singer, W. (2001). Dynamic predictions: oscillations and synchrony in top-down processing. *Nature Reviews Neuroscience*, 2 (10), 704-716. [10.1038/35094565](https://doi.org/10.1038/35094565)
- Engen, H. G. & Singer, T. (2013). Empathy circuits. *Current Opinion in Neurobiology*, 23 (3), 275-282. [10.1016/j.conb.2012.11.003](https://doi.org/10.1016/j.conb.2012.11.003)
- Felleman, D. J. & van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1 (1), 1-47. [10.1093/cercor/1.1.1](https://doi.org/10.1093/cercor/1.1.1)
- Fisch, L., Privman, E., Ramot, M., Harel, M., Nir, Y., Kipervasser, S., Andelman, F., Neufeld, M.Y., Kramer,

- U., Fried, I. & Malach, R. (2009). Neural „ignition“: Enhanced activation linked to perceptual awareness in human ventral stream visual cortex. *Neuron*, 64 (4), 562-574. [10.1016/j.neuron.2009.11.001](https://doi.org/10.1016/j.neuron.2009.11.001)
- Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C. & Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences of the USA*, 102 (27), 9673-9678. [10.1073/pnas.0504136102](https://doi.org/10.1073/pnas.0504136102)
- Friederici, A. D. & Gierhan, S. M. E. (2013). The language network. *Current Opinion in Neurobiology*, 23 (2), 250-254. [10.1016/j.conb.2012.10.002](https://doi.org/10.1016/j.conb.2012.10.002)
- Fries, P. (2005). A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends in Cognitive Sciences*, 9 (10), 474-480. [10.1016/j.tics.2005.08.011](https://doi.org/10.1016/j.tics.2005.08.011)
- (2009). Neuronal gamma-band synchronization as a fundamental process in cortical computation. *Annual Review of Neuroscience*, 32, 209-224. [10.1146/annurev.neuro.051508.135603](https://doi.org/10.1146/annurev.neuro.051508.135603)
- Fries, P., Roelfsema, P. R., Engel, A. K., König, P. & Singer, W. (1997). Synchronisation of oscillatory responses in visual cortex correlates with perception in interocular rivalry. *Proceedings of the National Academy of Sciences of the USA*, 101 (35), 13050-13055.
- Fries, P., Neuenschwander, S., Engel, A. K., Goebel, R. & Singer, W. (2001a). Rapid feature selective neuronal synchronization through correlated latency shifting. *Nature Neuroscience*, 4 (2), 194-200. [10.1038/84032](https://doi.org/10.1038/84032)
- Fries, P., Reynolds, J. H., Rorie, A. E. & Desimone, R. (2001b). Modulation of oscillatory neuronal synchronization by selective visual attention. *Science*, 291 (5508), 1560-1563. [10.1126/science.1055465](https://doi.org/10.1126/science.1055465)
- Fries, P., Schröder, J. H., Singer, W. & Engel, A. K. (2001c). Conditions of perceptual selection and suppression during interocular rivalry in strabismic and normal cats. *Vision Research*, 41 (6), 771-783. [10.1016/S0042-6989\(00\)00299-6](https://doi.org/10.1016/S0042-6989(00)00299-6)
- Gaillard, R., Dehaene, S., Adam, C., Clemenceau, S., Hasboun, D., Baulac, M., Cohen, L. & Naccache, L. (2009). Converging intracranial markers of conscious access. *PLOS Biology*, 7 (3), e1000061. [10.1371/journal.pbio.1000061](https://doi.org/10.1371/journal.pbio.1000061)
- Genc, E., Bergmann, J., Singer, W. & Koler, A. (2014). Surface area of early visual cortex predicts individual speed of travelling waves during binocular rivalry (forthcoming). *Cerebral Cortex*. [10.1093/cercor/bht342](https://doi.org/10.1093/cercor/bht342)
- Gray, C. M., König, P., Engel, A. K. & Singer, W. (1989). Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature*, 338, 334-337. [10.1038/338334a0](https://doi.org/10.1038/338334a0)
- Gray, C. M. & Singer, W. (1989). Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. *Proceedings of the National Academy of Sciences of the USA*, 86 (5), 1698-1702. [10.1073/pnas.86.5.1698](https://doi.org/10.1073/pnas.86.5.1698)
- Gross, J., Schmitz, F., Schnitzler, I., Kessler, K., Shaprio, K., Hommel, B. & Schnitzler, A. (2004). Modulation of long-range neural synchrony reflects temporal limitations of visual attention in humans. *Proceedings of the National Academy of Sciences of the USA*, 101 (35), 13050-13055. [10.1073/pnas.0404944101](https://doi.org/10.1073/pnas.0404944101)
- Hameroff, S. (2006). Consciousness, neurobiology and quantum mechanics. In J. Tuszynski (Ed.) *The Emerging Physics of Consciousness* (pp. 193-253). Berlin, GER: SpringerVerlag.
- Han, X., Chow, B. Y., Zhou, H., Klapoetke, N. C., Chuong, A. S., Rajimehr, R., Yang, A., Baratta, M. V., Winkle, J., Desimone, R. & Boyden, E. S. (2011). A high-light sensitivity optical neural silencer: Development and application to optogenetic control of non-human primate cortex. *Frontiers in Systems Neuroscience*, 5 (18), 1-8. [10.3389/fnsys.2011.00018](https://doi.org/10.3389/fnsys.2011.00018)
- Haynes, J.-D., Deichmann, R. & Rees, G. (2005). Eye-specific effects of binocular rivalry in the human lateral geniculate nucleus. *Nature*, 438, 496-499. [10.1038/nm1537](https://doi.org/10.1038/nm1537)
- Hipp, J. F., Engel, A. K. & Siegel, M. (2011). Oscillatory synchronization in large-scale cortical networks predicts perception. *Neuron*, 69 (2), 387-396. [10.1016/j.neuron.2010.12.027](https://doi.org/10.1016/j.neuron.2010.12.027)
- Hipp, J. F., Hawellek, D. J., Corbetta, M., Siegel, M. & Engel, A. K. (2012). Large-scale cortical correlation structure of spontaneous oscillatory activity. *Nature Neuroscience*, 15 (6), 884-890. [10.1038/nn.3101](https://doi.org/10.1038/nn.3101)
- Hodzic, A., Kaas, A., Muckli, L., Stirn, A. & Singer, W. (2009). Cortical responses to invisible objects in the human dorsal and ventral pathways. *NeuroImage*, 45 (4), 1264-1271. [10.1016/j.neuroimage.2009.01.027](https://doi.org/10.1016/j.neuroimage.2009.01.027)
- Houweling, A. R. & Brecht, M. (2008). Behavioural report of single neuron stimulation in somatosensory cortex. *Nature*, 451, 65-68. [10.1038/nature06447](https://doi.org/10.1038/nature06447)
- Iurilli, G., Ghezzi, D., Olcese, U., Lassi, G., Nazzaro, C., Tonini, R., Tucci, V., Benfenati, F. & Medini, P. (2012). Sound-driven synaptic inhibition in primary visual cortex. *Neuron*, 73 (3), 814-828. [10.1016/j.neuron.2011.12.026](https://doi.org/10.1016/j.neuron.2011.12.026)

- Jensen, O. & Tesche, C. D. (2002). Frontal theta activity in humans increases with memory load in a working memory task. *European Journal of Neuroscience*, 15 (8), 1395-1399. [10.1046/j.1460-9568.2002.01975.x](https://doi.org/10.1046/j.1460-9568.2002.01975.x)
- Kenet, T., Bibitchkov, D., Tsodyks, M., Grinvald, A. & Arieli, A. (2003). Spontaneously emerging cortical representations of visual attributes. *Nature*, 425, 954-956. [10.1038/nature02078](https://doi.org/10.1038/nature02078)
- Lau, H. C. & Passingham, R. E. (2007). Unconscious activation of the cognitive control system in the human prefrontal cortex. *The Journal of Neuroscience*, 27 (21), 5805-5811. [10.1523/JNEUROSCI.4335-06.2007](https://doi.org/10.1523/JNEUROSCI.4335-06.2007)
- Logothetis, N. K., Leopold, D. A. & Sheinberg, D. L. (1996). What is rivalling during binocular rivalry? *Nature*, 380, 621-624. [10.1038/380621a0](https://doi.org/10.1038/380621a0)
- Lukoševičius, M. & Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3 (3), 127-149. [10.1016/j.cosrev.2009.03.005](https://doi.org/10.1016/j.cosrev.2009.03.005)
- Malach, R. (2007). The measurement problem in human consciousness research. *Behavioral and Brain Sciences*, 30 (5-6), 516-517. [10.1017/S0140525X0700297X](https://doi.org/10.1017/S0140525X0700297X)
- Markov, N. T., Ercsey-Ravasz, M., Lamy, C., Ribeiro Gomes, A. R., Magrou, L., Misery, P., Giroud, P., Barone, P., Dehay, C., Toroczka, Z., Knoblauch, K., van Essen, D. C. & Kennedy, H. (2013). The role of long-range connections on the specificity of the macaque interareal cortical network. *Proceedings of the National Academy of Sciences of the USA*, 110 (13), 5187-5192. [10.1073/pnas.1218972110](https://doi.org/10.1073/pnas.1218972110)
- Markov, N. T. & Kennedy, H. (2013). The importance of being hierarchical. *Current Opinion in Neurobiology*, 23 (2), 187-194. [10.1016/j.conb.2012.12.008](https://doi.org/10.1016/j.conb.2012.12.008)
- Melloni, L., Molina, C., Pena, M., Torres, D., Singer, W. & Rodriguez, E. (2007). Synchronization of neural activity across cortical areas correlates with conscious perception. *The Journal of Neuroscience*, 27 (11), 2858-2865. [10.1523/JNEUROSCI.4623-06.2007](https://doi.org/10.1523/JNEUROSCI.4623-06.2007)
- Melloni, L. & Rodriguez, E. (2007). Non-perceived stimuli elicit global but not large-scale neural synchrony. *Perception*, 36
- Melloni, L. & Singer, W. (2011). The explanatory gap in neuroscience. *Pontificiae Academiae Scientiarum Acta*, 21, 61-73.
- Metzinger, T. (Ed.) (2000). *Neural correlates of consciousness: Empirical and conceptual questions*. Cambridge, MA: MIT Press.
- Muckli, L. & Petro, L. S. (2013). Network interactions: non-geniculate input to V1. *Current Opinion in Neurobiology*, 23 (2), 195-201. [10.1016/j.conb.2013.01.0](https://doi.org/10.1016/j.conb.2013.01.0)
- Nakatani, C., Ito, J., Nikolaev, A. R., Gong, P. & van Leeuwen, C. (2005). Phase synchronization analysis of EEG during attentional blink. *Journal of Cognitive Neuroscience*, 17 (12), 1969-1979. [10.1162/089892905775008706](https://doi.org/10.1162/089892905775008706)
- Penrose, R. (1994). *Shadows of the mind: A search for the missing science of consciousness*. New York, NY: Oxford University Press.
- Power, J. D. & Petersen, S. E. (2013). Control-related systems in the human brain. *Current Opinion in Neurobiology*, 23 (2), 223-228. [10.1016/j.conb.2012.12.009](https://doi.org/10.1016/j.conb.2012.12.009)
- Raichle, M. E. (2011). The restless brain. *Brain Connectivity*, 1 (1), 3-12. [10.1089/brain.2011.0019](https://doi.org/10.1089/brain.2011.0019)
- Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A. & Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences of the USA*, 98 (2), 676-682. [10.1073/pnas.98.2.676](https://doi.org/10.1073/pnas.98.2.676)
- Rodriguez, E., George, N., Lachaux, J. P., Martinerie, J., Renault, B. & Varela, F. J. (1999). Perception's shadow: Long distance synchronization of human brain activity. *Nature*, 397 (6718), 430-433. [10.1073/pnas.98.2.676](https://doi.org/10.1073/pnas.98.2.676)
- Roopun, A. K., Kramer, M. A., Carracedo, L. M., Kaiser, M., Davies, C. H., Traub, R. D., Kopell, N. J. & Whittington, M. A. (2008). Temporal interactions between cortical rhythms. *Frontiers in Neuroscience*, 2 (2), 145-154. [10.1126/science.1099745](https://doi.org/10.1126/science.1099745)
- Salazar, R. F., Dotson, N. M., Bressler, S. L. & Gray, C. M. (2012). Content specific fronto-parietal synchronization during visual working memory. *Science*, 338 (6110), 1097-1100. [10.1126/science.1224000](https://doi.org/10.1126/science.1224000)
- Salzman, C. D., Murasugi, C. M., Britten, K. H. & Newsome, W. T. (1992). Microstimulation in visual area MT: Effects on direction discrimination performance. *The Journal of Neuroscience*, 12 (6), 2331-2355.
- Schack, B., Klimesch, W. & Sauseng, P. (2005). Phase synchronisation between theta and upper alpha oscillations in a working memory task. *International Journal of Psychophysiology*, 57 (2), 105-114. [10.1016/j.ijpsycho.2005.03.016](https://doi.org/10.1016/j.ijpsycho.2005.03.016)
- Schwarz, C. & Bolz, J. (1991). Functional specificity of the long-range horizontal connections in cat visual cortex: A cross-correlation study. *The Journal of Neuroscience*, 11 (10), 2995-3007.
- Searle, J. R. (1997). *The mystery of consciousness*. London, UK: Granta Books.

- Sherrington, C. S. (1906). *The integrative action of the nervous system*. New York, NY: Charles Scribner's Sons.
- Shew, W. L., Yang, H., Petermann, T., Roy, R. & Plenz, D. (2009). Neuronal avalanches imply maximum dynamic range in cortical networks at criticality. *The Journal of Neuroscience*, 29 (49), 15595-15600. [10.1523/JNEUROSCI.3864-09.2009](https://doi.org/10.1523/JNEUROSCI.3864-09.2009)
- Silver, M. A. & Logothetis, N. K. (2004). Grouping and segmentation in binocular rivalry. *Vision Research*, 44 (14), 1675-1692. [10.1016/j.visres.2003.12.008](https://doi.org/10.1016/j.visres.2003.12.008)
- Singer, W. (1999). Neuronal synchrony: A versatile code for the definition of relations? *Neuron*, 24 (1), 49-65. [10.1016/S0896-6273\(00\)80821-1](https://doi.org/10.1016/S0896-6273(00)80821-1)
- (2009). Genetic and epigenetic shaping of cognition – prerequisites of cultural evolution. In W. Arber, N. Cabibbo & M. Sánchez Sorondo (Eds.) *The proceedings of the plenary session on scientific insights into the evolution of the universe and of life. 31 October – 4 November 2008, vol Pontificiae Academiae Scientiarum Acta* (pp. 337-347). Vatican City, VA: Pontificia Academia Scientiarum.
- (2010). Neocortical rhythms. An overview. In C. von der Malsburg, W. A. Phillips & W. Singer (Eds.) *Dynamic coordination in the brain. From neurons to mind* (pp. 159-168). Cambridge, MA: MIT Press.
- (2013). Cortical dynamics revisited. *Trends in Cognitive Sciences*, 17 (12), 616-626. [10.1016/j.tics.2013.09.006](https://doi.org/10.1016/j.tics.2013.09.006)
- Sporns, O. (2013). Network attributes for segregation and integration in the human brain. *Current Opinion in Neurobiology*, 23 (2), 162-171. [10.1016/j.conb.2012.11.015](https://doi.org/10.1016/j.conb.2012.11.015)
- Srinivasan, R., Russell, D. P., Edelman, G. M. & Tononi, G. (1999). Increased synchronization of neuromagnetic responses during conscious perception. *The Journal of Neuroscience*, 19 (13), 5435-5448.
- Stokes, M. G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D. & Duncan, J. (2013). Dynamic coding for cognitive control in prefrontal cortex. *Neuron*, 78 (2), 364-375. [10.1016/j.neuron.2013.01.039](https://doi.org/10.1016/j.neuron.2013.01.039)
- Uhlhaas, P. J., Pipa, G., Lima, B., Melloni, L., Neuenschwander, S., Nikolic, D. & Singer, W. (2009). Neuronal synchrony in cortical networks: history, concept and current status. *Frontiers in Integrative Neuroscience*, 3 (17), 1-19. [10.3389/neuro.07.017.2009](https://doi.org/10.3389/neuro.07.017.2009)
- Van den Heuvel, M. P. & Sporns, O. (2011). Rich club organization of the human connectome. *The Journal of Neuroscience*, 31 (44), 15775-15786. [10.1523/JNEUROSCI.3539-11.2011](https://doi.org/10.1523/JNEUROSCI.3539-11.2011)
- (2013). Network hubs in the human brain. *Trends in Cognitive Sciences*, 17 (12), 683-696. [10.1016/j.tics.2013.09.012](https://doi.org/10.1016/j.tics.2013.09.012)
- Van Gaal, S., Ridderinkhof, K. R., Fahrenfort, J. J., Scholte, H. S. & Lamme, V. A. (2008). Frontal cortex mediates unconsciously triggered inhibitory control. *The Journal of Neuroscience*, 28 (32), 8053-8062. [10.1523/JNEUROSCI.1278-08.2008](https://doi.org/10.1523/JNEUROSCI.1278-08.2008)
- Varela, F., Lachaux, J. P., Rodriguez, E. & Martinerie, J. (2001). The brainweb: Phase synchronisation and large-scale integration. *Nature Reviews Neuroscience*, 2 (4), 229-239. [10.1038/35067550](https://doi.org/10.1038/35067550)
- Vinje, W. E. & Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287 (5456), 1273-1276. [10.1126/science.287.5456.1273](https://doi.org/10.1126/science.287.5456.1273)
- von der Malsburg, C., Phillips, W. A. & Singer, W. (2010). *Dynamic coordination in the brain. From neurons to mind*. Cambridge, MA: MIT Press.
- Zeki, S. (2001). Localization and globalization in conscious vision. *Annual Review of Neuroscience*, 24, 57-86. [10.1146/annurev.neuro.24.1.57](https://doi.org/10.1146/annurev.neuro.24.1.57)

It's Not Just About the Contents: Searching for a Neural Correlate of a State of Consciousness

A Commentary on Wolf Singer

Valdas Noreika

Global gamma band synchronisation is perhaps the most extensively studied candidate for a Neural Correlate of Consciousness (NCC). Yet despite numerous studies confirming its association with consciousness, it seems to be neither sufficient nor necessary for the presence of all subjective experiences. Analysis of gamma synchronisation studies suggests that it is a correlate of the initial binding of expected, attended, task-dependent contents of consciousness, whereas task-irrelevant contents do not seem to require gamma synchronisation. While discovery of such a content-related NCC is a remarkable achievement for the neurophysiological research of consciousness, it does not fully explain some of the fundamental structural properties of consciousness, namely the temporal and spatial integration of all available experiences into a coherent stream of consciousness. As an alternative, instead of focusing solely on the selected contents of consciousness, the neural mechanisms of the fundamental properties of consciousness could be studied by contrasting states of (un)consciousness. Recent research into the states of consciousness suggests that, for instance, informational complexity is a highly sensitive predictor of the presence of consciousness, possibly reflecting background structural properties of the unity of subjective experiences. As a limiting factor, though, such a state-related NCC does not seem to reflect the phenomenal diversity of the contents of consciousness. Arguably, these limitations could be overcome by devising experimental setups that would simultaneously probe the neural correlates of the contents and the state of consciousness.

Keywords

Contents of consciousness | Gamma band synchronisation | Neural correlate of consciousness (NCC) | Neural correlates of consciousness | Nonconscious states | Spatial binding | State of consciousness | Stream of consciousness | Temporal binding | Unconscious states

1 Introduction

Even though the search for the neural correlates of consciousness is still an unresolved challenge of astonishing complexity (Crick 1994), the continuous efforts to crack the mystery are not expended in vain. Each year brings an increasing number of cognitive neuroscientific studies that reveal yet another piece of the puzzle of the neural basis of subjective experience. However,

it often seems that individual findings are too diverse and sparse to form a coherent picture. In addition to the fundamental problem of the binding of conscious experiences (Singer 2001), we increasingly face the problem of how to bind the findings of consciousness-related studies. The present target paper by Prof. Wolf Singer serves such a discovery-binding function, bring-

Commentator

Valdas Noreika

valdas.noreika@mrc-cbu.cam.ac.uk

Medical Research Council
Cambridge, England

Target Author

Wolf Singer

w.singer@brain.mpg.de

Max Planck Institute for Brain
Research (MPI)
Frankfurt a. M., Germany

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

ing very diverse findings into a unified picture of how the neural correlates with the subjective.

In an impressively erudite manner, [Singer \(this collection\)](#) integrates a very broad range of anatomical and functional findings of the organisational principles of the brain, concluding that the high-level cognitive functions are supported by densely coupled, recurrent neural networks, interacting under the principles of non-linear dynamics. In the proposed framework, perception is treated as an active process, whose self-organisation is initially determined by genes, and later modified by post-natal development, learning, social interactions, and cultural influences. At the neuronal networks level, high-level integration and communication are achieved through synchronisation of oscillations in different electroencephalography (EEG) frequency bands, the most notable of which is gamma band ($>30\text{Hz}$) synchronisation ([Engel et al. 1999](#)). Given that an association between the widespread gamma-band synchronisation and conscious awareness is found in rather different experimental paradigms, such as visual masking ([Melloni et al. 2007](#)), binocular rivalry ([Doeburg et al. 2009](#)), and attentional blink ([Gross et al. 2004](#)), gamma synchronisation is often regarded as the main NCC ([Singer this collection](#)).

Yet the candidature of gamma synchronisation as the correlate of consciousness is challenged by some findings from research into the behavioural states of the brain. If gamma-range activity correlates with consciousness, it should diminish when consciousness ceases. Contrary to this, gamma band activity seems to increase rather than decrease in response to certain general anaesthetics, such as ketamine ([Steriade et al. 1996](#)). Furthermore, gamma synchronisation seems to be absent in some conscious brain states. For instance, it has been reported that large-scale neocortical gamma-band coherence is virtually absent during rapid eye movement (REM) sleep in cats ([Castro et al. 2013](#)), a state typically marked by the most intense dreaming in humans ([Hobson et al. 2000](#)) as well as in felids ([Jouvet 1979](#)).

The target paper briefly mentions neural mechanisms supporting overall brain states, but

dismisses them as modulatory systems that are too general to be considered the NCC ([Singer this collection](#)). In the following, I will argue that the relation between the neural mechanisms of the contents and states of consciousness is not straightforward, and that the puzzle of the neural mechanisms of consciousness cannot be completed without studying the neural mechanisms of conscious states.

2 Contents vs. states of consciousness

An important distinction in consciousness research is that between the contents and a state of consciousness ([Chalmers 2000](#)). The concept of the contents of consciousness refers to individual subjective experiences that occur in phenomenal consciousness, such as reading a word or hearing birdsong, and as such they are sometimes referred to as the phenomenal contents of consciousness ([Revonsuo 2006](#)). Most neural experiments on consciousness, especially in the dominant field of visual awareness studies, are concerned with the neural basis of such specific contents of consciousness, i.e., they select one or two subjective experiences within an overall stream of consciousness. In this type of experiment, participants may be presented with stimuli close to their perceptual threshold ([Del Cul et al. 2009](#)) or they may be instructed to observe ambiguous stimuli that may lead to perception of several alternating contents of consciousness ([Kornmeier & Bach 2012](#)). The brain responses are then contrasted between trials that differ in awareness of these stimuli. Notably, while participants in such experiments report being unaware of some contents of consciousness, they still maintain awareness of other experiences: such as seeing the edges of a computer screen, hearing the background noise of the Magnetic Resonance Imaging (MRI) scanner, or letting their thoughts wander away from the experimental task. Typically, such experiences are ignored as task-irrelevant, and consequently the so-called “unaware” or “unconscious” trials still bear very rich phenomenology.

Contrary to the selective contents of consciousness, the concept of the state of consciousness refers to an overall pattern of subjective

psychological functioning that includes the totality of phenomenal contents of consciousness (Rosenthal 1986; Tart 1972). In addition to the relaxed waking state of consciousness in a healthy volunteer, which could be also regarded as a *baseline state*, altered, unconscious, and non-conscious states can be distinguished. In *altered states of consciousness*, such as dreaming or Lysergic Acid Diethylamide (LSD) psychomodulation, subjective experiences may undergo various perceptual and cognitive alterations, the neural basis of which can be studied by contrasting them with a baseline state of consciousness, e.g., by comparing brain activity before and after hallucinogen intake (Carhart-Harris et al. 2012). Given that there is no widely accepted definition and criterion for an altered state of consciousness (Móro 2010; Revonsuo et al. 2009), a rather common approach is to describe, classify, and study states that are traditionally called *altered*, avoiding a single definition that would grasp the core of all altered states of consciousness (Vaitl et al. 2005).

Contrary to the baseline and altered states of consciousness, unconscious states are deprived of subjective experiences, but they may still maintain the potential to become conscious. For instance, an unconscious state of dreamless sleep may turn into a conscious sleep once a sleeping participant begins to dream (for an alternative interpretation of dreamless sleep, see Thompson this collection). Finally, non-conscious states are those completely deprived of a capacity to support phenomenal consciousness, such as an irreversible coma. In clinical neuroscience, the most extensively studied contrast between a pathological altered state of consciousness and an unconscious or non-conscious state is a comparison between minimally conscious and vegetative state patients (Sitt et al. 2014). When states of (un)consciousness are contrasted, neural representations of specific experiences are typically ignored, making it difficult or even impossible to assess the phenomenal specificity of findings, e.g., if participants were aware of particular external stimuli or what internally generated experiences they had. Nevertheless, research into these states may re-

veal neural patterns that are common to all subjective experiences without individuating them.

It is possible that these two lines of research may eventually reveal rather different, if not independent, NCC systems: a neural correlate of the state of consciousness and a neural correlate of the contents of consciousness (Chalmers 2000). If these exist, any neuroscientific program of consciousness research would be incomplete without searching for a state NCC. Furthermore, even if a separate state NCC did not exist, there is currently no evidence for this, and thus NCC research is incomplete if it does not investigate this possibility. This view is often dismissed on the basis that some of the most plausible candidates for a state NCC, such as the brainstem reticular formation (Merker 2007; Parvizi & Damasio 2001), are rather low-level neural systems, whereas converging evidence shows that consciousness is a cortical process (Singer this collection). Furthermore, it could be argued that a conscious state may be nothing more than the sum of individual experiences, in which case revealing the NCC of specific contents would automatically explain the state NCC. Yet a brief analysis of the fundamental structural properties of consciousness—see the following section—shows that the necessary and sufficient NCC cannot be revealed by an exclusive focus on the contents of consciousness. Notably, the arguments presented in this commentary will be confined to the biological nature of human and animal consciousness, and as such they are not applicable to the problem of machine, extraterrestrial, or silicon-brain consciousness.

3 Unity as the fundamental property of the stream of consciousness

Given that subjective experiences can accompany almost any sensory, cognitive, emotional, and behavioural function of the brain, phenomenal consciousness turns out to be an extremely complex and multi-dimensional process. Nevertheless, introspection shows that despite its qualitative richness, phenomenal consciousness appears to us as a unified and coherent model

of the external and internal environment (Bayne 2010; Revonsuo 2006). The continuity in the diversity of subjective experiences is famously referred to as the stream of consciousness (James 1890). This metaphor points to the unification of experiences occurring at different points in time and space, which is achieved through temporal and spatial binding.

At the cognitive level of description, temporal binding, i.e., integration of subjective experiences over time, is realised through the perception of simultaneity, duration, and successiveness (Kiverstein 2010; Pöppel 1997). Perception of simultaneity may integrate several experiences, e.g., seeing a cat and hearing a birdsong in the park, as occurring at the same time. Perception of duration of selected experiences may extend them in time, e.g., the birdsong might seem to last for a certain period of time. Finally, perception of successiveness may signal the end of one temporally-extended experience, and the beginning of another one, e.g., as the cat reaches the bush and the birdsong ceases, we may notice a cone under the bush. Notably, the change does not typically involve all experiences, and as we are aware of some changing contents, some other experiences continue to endure in time, e.g., we still see the same bush. In addition to the timing-specific functions, temporal binding seems to depend on the iconic memory that contains the just-experienced contents of consciousness, and on the anticipation of subsequent ones, forming the temporally-extended phenomenal experience of now, sometimes referred to as the *specious present* (Dainton 2006; Kelly 1882). Temporal extension of subjective experiences have a simple, yet very important and often overlooked implication for NCC research: if there is a single neural mechanism generating phenomenal consciousness, it should be present as long as we are conscious of at least one single content. Given that our experiences do not cease at a fixed rate, i.e., some are shorter and some longer, and that the change does not happen abruptly for all experiences at once, the NCC should persist for the duration of the stream of consciousness. Thus, a temporary-confined correlate of awareness, such as a negative Event Related Potential (ERP)

waveform briefly peaking at about 200ms from the onset of visual stimuli (Railo et al. 2011), cannot be a sufficient correlate of consciousness, as awareness of visual contents lasts for a much longer period of time.

Spatial binding, i.e., integration of subjective experiences in space, is realised through several complimentary processes, through which each subjective experience occupies a specific location in relation to other experiences, which is sometimes referred to as *location binding* (Treisman 1996). In the baseline state of consciousness, one experience never occurs in isolation from other experiences; and when some experiences cease, we do not experience emptiness, because other experiences fill in their place. Furthermore, individual experiences are spatially integrated not only with respect to each other in phenomenally external space, but also with respect to the common egocentric reference point (Revonsuo 2006). The reference point is typically located in the phenomenal head or chest, and all other experiences are realised in the space as taking a certain distance and angle from this point. While typical phenomenological analysis of 3D space considers visual and auditory experiences, it has recently been shown that emotions and feelings are also experienced as taking certain location with respect to our body parts (Nummenmaa et al. 2014). For instance, anger is overrepresented in hands and arms when compared to sadness. Arguably, even thoughts, which are often regarded as non-spatial entities (Clarke 1995), are usually experienced as occurring within the head rather than somewhere else. Given that the whole phenomenal space is bound together, the NCC should also represent awareness of the whole space rather than, for instance, selected regions on the computer screen. That is, a promising candidate for an NCC should not cease when a specific experience vanishes as long as the spatial and temporal unity of the stream of consciousness is maintained.

So, what type of neural processes should we be looking for when searching for the NCC? If we take the unity of consciousness seriously, we should be looking for a neural process that steadily represents the whole phenomenal space,

and sustains its activity over periods of time longer than the existence of a single experience. Arguably, the neural correlate of unified consciousness cannot be discovered by studying and contrasting only isolated contents of consciousness, as the unity of spatiotemporal interactions simply cannot be derived from solitary experiences. Thus, while continuing to search for the neural mechanisms of the contents of consciousness, the NCC program should be extended by carrying out systematic contrasts between unconscious, baseline, and/or altered states, which would consider the whole stream of consciousness. A possible objection to this proposal is that the unity of consciousness is not fundamental in the strong form of fundamentalism, i.e., some forms of consciousness may still exist despite the possible disintegration of its unity, which seems to happen in states like schizophrenia, sleep onset, or a minimally conscious state. For instance, the stream of consciousness may occasionally undergo a sudden, unpredictable alteration in terms of inner speech and imagery (Noreika et al. 2014). Nevertheless, if one aims to explain the neural mechanisms of normal waking consciousness, the unity thesis, with its NCC-related implications, cannot be ignored. With these considerations in mind, let us examine now the proposal of global gamma synchronisation as the NCC (Singer this collection).

4 Is gamma band synchrony sufficient and necessary for consciousness?

A sufficient and necessary NCC (or perhaps a set of NCCs) should be generic enough to cover all conscious contents and states, and should also be specific enough to cover only conscious contents and states. Notably, gamma synchronisation does not meet the second specificity requirement, as it can be associated with almost any perceptual and cognitive function that depends on the formation of temporary associations of distributed neuronal networks. Among numerous cases, increased gamma-band synchronisation is found to be associated with such tasks as perceptual learning (Gruber et al. 2002), self-paced movement (Pfurtscheller et al.

2003), mental rotation (Bhattacharya et al. 2001), viewing of unpleasant stimuli (Martini et al. 2012), deductive reasoning (Zhang et al. 2014), auditory attention control (Doesburg et al. 2012), face integration (Kottlow et al. 2012), and memory encoding and retrieval (Osipova et al. 2006). Gamma-band synchronisation thus seems to be a generic process that contributes to complex cortical computations involved in most if not all of the higher cognitive functions (Fries 2009).

Singer (this collection) proposes that only global, widespread synchronisation of gamma oscillations is associated with consciousness, whereas local, spatially-restricted synchronisation is not necessarily related to conscious awareness. This might refute studies reporting local synchrony; however, some of the above-mentioned studies found increased global gamma-synchrony when observing unpleasant stimuli (Martini et al. 2012) or carrying out a mental rotation task (Bhattacharya et al. 2001). It could be argued, though, that gamma synchronisation is present in these experiments as a correlate of task-related subjective experiences, such as awareness of memory retrieval. In fact, even though most of these studies did not even mention consciousness or awareness, their participants were not unconscious, and gamma synchronisation could have been associated with the task-dependent subjective experiences. Yet, this line of reasoning is challenged by the simple fact that participants remained conscious in all contrast conditions throughout the experiments. Why would consciousness-related gamma synchronisation increase in some, but not other conditions? This leads us to the question of what exactly gamma synchrony correlates with in studies that specifically manipulate awareness? Let us take a closer look at two key studies, also examined by Singer (this collection).

Melloni et al. (2007) presented pairs of words and asked participants to report on whether both words were the same. Visibility of the first word was manipulated by adjusting the luminance level of the forward and backward masks, which rendered the words visible only in some of the trials. Global gamma-phase synchronisation between the fronto-centro-parietal

electrodes was observed within the 40–182ms time-window after the presentation of the first word only in visible trials, which coincides with the time when conscious perception of the words is expected to emerge. In the latter time-windows, visible words were marked by more localised gamma synchronisation, higher P300 amplitude, and higher amplitude of frontal theta oscillations than invisible words. These findings confirmed that gamma synchronisation is a correlate of visual-semantic awareness, and showed that other electrophysiological processes may also correlate with consciousness.

Doesburg et al. (2009) investigated the role of gamma-phase synchronisation in conscious awareness using a binocular rivalry paradigm, in which a different visual stimulus is presented to each eye. Instead of seeing both stimuli at the same time, people report perceiving only one of the stimuli that continues switching in time. An increase in the gamma-band synchronisation over the fronto-parietal regions was observed in the 600–540ms and 280–220ms time-windows before responses indicating a perceptual switch. Assuming that reaction time was about 250ms, the synchronisation increase coincided with a new percept reaching awareness. Interestingly, gamma synchronisation oscillated at the theta rhythm, suggesting a cross-frequency interaction.

In both of these experiments (Doesburg et al. 2009; Melloni et al. 2007), gamma synchrony peaked around the time when participants began experiencing a new content of consciousness, following which Singer (this collection) draws the well-justified conclusion that gamma synchronisation is associated with a transfer of the new contents into awareness. Given that increased synchronisation may reflect the neural and phenomenal binding required for the fundamental unity of consciousness to emerge, it seems to be an ideal candidate for the NCC. Yet the duration of increased synchronisation is relatively brief and seems to last a much shorter time than the awareness of stimuli. For instance, P300 distinguished visible and invisible words around 300ms post-stimulus, whereas gamma-band synchronisation became local during this time-window (Melloni et al. 2007). Such

brevity of synchronisation suggests that it is involved only in the initial binding of the new contents of consciousness, while a further maintenance of these contents is supported by other neural mechanisms, in particular theta oscillations (Doesburg et al. 2009; Melloni et al. 2007; Singer this collection). Given that the global gamma synchronisation correlates with a spatially- and temporally-local change in the stream of consciousness, its association with an overall unity of consciousness is uncertain and, at least currently, it cannot be accounted as the only or even as the major NCC. If it were such, it would not cease as long as the participant were aware of a particular content of consciousness.

Furthermore, gamma synchrony does not seem to increase in response to each of the new contents of consciousness. In each trial, Melloni et al. (2007) presented a series of stimuli, including a fixation cross, a masking noise, a target word, and a blank screen. Each of these stimuli should have entered consciousness, and even when the target word was unreadable, participants should have perceived something, e.g., an unreadable word, incoherent letters, or a flashing mask. However, the global gamma synchronisation increased only in response to perceived visible words, suggesting that it is a correlate of the initial binding of a selected, expected, attended, coherent, task-relevant content of consciousness. As such, in addition to the lack of specificity, gamma synchrony does not seem to be generic enough to cover all the different contents of consciousness, even within the paradigmatic visual modality. Thus, it seems that the global gamma synchronisation is neither necessary nor sufficient for consciousness to emerge, as subjective experiences may exist without gamma synchronisation, and even when synchronisation is involved in the generation of awareness, other neural processes are needed to maintain its presence.

As discussed in the previous section, the unity of consciousness emerges from the interaction of all experiences available at a time. Gamma synchronisation cannot account for the unity of a state of consciousness, simply because it is involved only in the generation of new task-

dependent contents, and it does not seem to bind these contents within the broader stream of consciousness. Arguably, instead of focusing on selected stimuli, we may be able to detect the neural correlates of the unity of consciousness by contrasting states of consciousness with unconsciousness, since such a contrast would consider the whole stream of phenomenal contents, including their structural unity.

5 Studying the contents and states of consciousness: Let's probe them together!

Perhaps the most powerful contrast conditions for studying a neural correlate of a state of consciousness are comparisons between wakefulness and slow-wave sleep, as well as between wakefulness and general anaesthesia. A notable clinical contrast is a comparison between vegetative state and minimally conscious state patients. Furthermore, new paradigms are available for comparing consciousness with unconsciousness when an overall physiological state of the brain is controlled, such as dreamless vs. dreamful non-rapid eye movement sleep (NREM sleep; Noreika et al. 2009; Siclari et al. 2013) or dreamless vs. dreamful anaesthesia (Noreika et al. 2011). Let us examine several exemplary papers that compare an overall stream of consciousness with its absence.

Sitt et al. (2014) studied auditory-evoked potentials and endogenous fluctuations of EEG signal in 75 vegetative state and 68 minimally-conscious patients. None of the studied evoked potentials (P1, MMN, P3a, P3b, CNV) were able to discriminate patient groups, indicating that task-dependent brain activity does not necessarily distinguish between conscious and unconscious states. Contrary to this, analyses of spontaneous EEG activity showed that unconscious patients had higher power of delta and lower power of theta and alpha oscillations, especially over parietal regions. Furthermore, EEG complexity indices derived from the compressibility of a sequence of data points indicated increased signal complexity over the parietal region in the minimally conscious patients compared to the vegetative state patients. Fi-

nally, electrode connectivity measures derived from information theory showed that vegetative-state patients had lower-weighted symbolic mutual information exchange in the range of theta and alpha oscillations than minimally-conscious patients. Interestingly, none of the EEG connectivity measures in the gamma frequency range, including phase lag index and imaginary coherence, could discriminate patient groups, coinciding with other independent observations that gamma synchrony does not necessarily differentiate conscious and unconscious states of the brain (Castro et al. 2013; Steriade et al. 1996).

The finding that the presence of consciousness is associated with an overall complexity of EEG signal and the magnitude of inter-electrode information exchange (Sitt et al. 2014) seems to support the information integration theory of consciousness (Tononi 2012), which predicts that consciousness depends on information complexity and integration in the system. The information integration theory was recently tested by Casali et al. (2013), who investigated the consciousness-related electrodynamics of the distributed cortical networks in a wide range of states of (un)consciousness, including wakefulness (eyes open, eyes closed), sleep (NREM sleep, REM sleep), anaesthesia (midazolam, xenon, propofol), and consciousness disorders (locked-in syndrome, minimally conscious state, patients who have emerged from a minimally conscious state, vegetative state). In a series of experiments, transcranial magnetic stimulation (TMS) pulses were delivered to different cortical sites, which perturbed spontaneous EEG activity (Massimini et al. 2010). Complexity of such TMS-induced EEG perturbations was then calculated, and its index successfully differentiated the states of consciousness and unconsciousness, even at the individual participant's level (Casali et al. 2013). As predicted, the presence of consciousness was associated with a higher level of information complexity.

In these and similar experiments, the contents of consciousness were not systematically manipulated or controlled for, and conscious participants probably underwent very diverse experiences. Consequently, the reported EEG

complexity as the NCC seems to be independent of particular phenomenal contents, and it may reflect some structural aspects of the whole stream of unified subjective experiences. It seems that phenomenal consciousness emerges in a state of the brain that is capable of generating the required level of information complexity and integration. As requested in the previous sections, such an NCC is stable in time and does not depend on an experience isolated from the rest of phenomenal space. Arguably, this type of study tackles the fundamental unity of consciousness much more directly than typical paradigms for studying the selected contents of consciousness. However, approaching one side of the bridge takes us further away from the other side, and the better characterization we have of the neural mechanisms of the state of consciousness, the less we can say about the neural mechanisms of particular contents of consciousness. For instance, the perturbational complexity index can differentiate conscious and unconscious states, but it is extremely insensitive when it comes to distinguishing between different contents of consciousness. For instance, the values of the complexity index did not systematically differ between the “eyes closed” and “eyes open” conditions in the standard waking state (Casali et al. 2013; Noreika 2014). Arguably, any NCC that cannot distinguish between experiences occurring in the “eyes closed” and “eyes open” conditions cannot be fully satisfactory, as the quality of subjective experiences is the core of the scientific problem of consciousness. Yet even though informational complexity does not reflect qualities of phenomenal contents, it is a promising candidate for an NCC of the background properties of consciousness that enable the emergence of subjective experiences and/or necessitate their structural unity.

We are thus left with studies of the contents NCC, such as focusing on the gamma synchrony, and studies of the state NCC, such as focusing on the information complexity. The first group of studies seems to explain the neural binding of concrete selected contents of consciousness, but it does not have a capacity to address the unity of consciousness. The second group seems to capture neural processes

involved in the whole stream of consciousness, but it ignores differentiation or phenomenal diversity of consciousness. Ideally, research into the NCC would combine both of these complementary approaches. Unfortunately, a systematic combination of the contents- and states-focused paradigms is almost never tested in cognitive neuroscientific studies of consciousness.

The combined *contents-states paradigm* would contrast baseline and altered states, or consciousness and unconsciousness, or the transition between the two, while participants carry out experiments that tackle the neural mechanisms of the contents of consciousness. For instance, one could study binocular rivalry while participants lose consciousness in response to an anaesthetic agent. This could, for instance, provide data to investigate how global gamma synchrony as a correlate of the binding and transfer of new contents to awareness depends on or interacts with a changing level of neuronal information complexity. Another promising avenue is research into awareness-related performance in the transition from wakefulness to sleep (Goupil & Bekinschtein 2011). In a recent attempt, Bareham et al. (2014) demonstrated that healthy individuals show neglect-like loss of awareness of the right side of their space in a drowsy state of consciousness. Thus, spatial awareness and unity seem to depend on the state of alertness, as defined by the relative amplitude of theta and alpha oscillations, which confirms that the contents are not wholly independent of the state. That is, despite the external physical stimuli and environment remaining stable, phenomenal contents may appear, disappear, or reorganise depending on the overall state of consciousness. More such studies are expected to be carried out in future, aiming to integrate the content NCC, the state NCC, and altered states of consciousness research programs under one unified framework of the content-state NCC research.

6 Conclusion

Global gamma-band synchronisation, research into which was largely triggered and continues to be advanced by Prof. Wolf Singer, is one of

the most promising NCCs. Synchronisation seems to increase in most cases when a new, task-dependent content of consciousness is formed. Yet a larger number of complications prevents its acceptance as the main NCC, namely: gamma synchronisation does not persist for as long as the contents of consciousness, some of the contents of consciousness emerge without gamma synchronisation being modulated, and, finally, gamma synchronisation may increase in unconscious or unresponsive states of mind. These complications show that gamma-band synchronisation cannot fully account for the existence of a unified stream of consciousness. Given that consciousness is integrated over cognitive time and space, a sufficient and necessary NCC should persist even when some but not all of the experiences cease to exist in time, or change their location. Nevertheless, even though gamma-band synchronisation seems to be neither necessary nor sufficient for all contents of consciousness to arise, it should be regarded as one of the NCCs specifically involved in the binding of new attended experiences. Future research may also develop more accurate characterization of gamma synchronisation, including its spatial scale, precision, and stability (Singer [this collection](#)), and certain forms of synchrony might be necessarily accompanied by consciousness; yet such evidence is not currently available.

Given that gamma synchronisation cannot be the only NCC, research efforts and resources should be distributed to search for the other NCCs, some of which might be responsible for the maintenance of already-bound single contents, and some of which might contribute to the unity of the whole stream of consciousness. Research paradigms should be developed that allow simultaneous manipulation and testing of both the contents and the states of (un)consciousness. Most likely, none of the discovered NCCs alone will be necessary and sufficient for all forms of subjective experiences to exist. How many of the neural correlates will be sufficient for the stream of consciousness to flow, and whether the sufficient ones will also be necessary, remains to be studied in future. For now, an exciting program of NCC research should

continue searching for the new avenues. Among various proposals, such as a focus on how social interactions and culture modulate neural networks supporting phenomenal contents (Singer [this collection](#)), the present one claims that it's not just about the contents, and that a state of consciousness deserves a treatment of its own.

Acknowledgements

My research is supported by a Wellcome Trust Biomedical Research Fellowship WT093811MA (awarded to Dr. Tristan Bekinschtein).

References

- Bareham, C. A., Manly, T., Pustovaya, O. V., Scott, S. K. & Bekinschtein, T. A. (2014). Losing the left side of the world: Rightward shift in human spatial attention with sleep onset. *Scientific Reports*, 4, 5092-5092. [10.1038/srep05092](https://doi.org/10.1038/srep05092)
- Bayne, T. (2010). *The Unity of Consciousness*. Oxford, UK: Oxford University Press.
- Bhattacharya, J., Petsche, H., Feldmann, U. & Rescher, B. (2001). EEG gamma-band phase synchronization between posterior and frontal cortex during mental rotation in humans. *Neuroscience Letters*, 311, 29-32. [10.1016/S0304-3940\(01\)02133-4](https://doi.org/10.1016/S0304-3940(01)02133-4)
- Carhart-Harris, R. L., Erritzoe, D., Williams, T., Stone, J. M., Reed, L. J., Colasanti, A., Tyacke, R. J., Leech, R., Malizia, A. L., Murphy, K., Hobden, P., Evans, J., Feilding, A., Wise, R. G. & Nutt, D. J. (2012). Neural correlates of the psychedelic state as determined by fMRI studies with psilocybin. *Proceedings of the National Academy of Sciences*, 109 (6), 2138-2143. [10.1016/S0304-3940\(01\)02133-4](https://doi.org/10.1016/S0304-3940(01)02133-4)
- Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., Casarotto, S., Bruno, M.-A., Laureys, S., Tononi, G. & Massimini, M. (2013). A theoretically based index of consciousness independent of sensory processing and behavior. *Science Translational Medicine*, 5 (198), 198ra105-198ra105. [10.1126/scitranslmed.3006294](https://doi.org/10.1126/scitranslmed.3006294)
- Castro, S., Falconi, A., Chase, M. H. & Torterolo, P. (2013). Coherent neocortical 40-Hz oscillations are not present during REM sleep. *European Journal of Neuroscience*, 37 (8), 1330-1339. [10.1111/ejn.12143](https://doi.org/10.1111/ejn.12143)
- Chalmers, D. J. (2000). What is a neural correlate of consciousness? In T. Metzinger (Ed.) *Neural Correlates of Consciousness: Empirical and Conceptual Questions* (pp. 31-63). Cambridge, MA: MIT Press.
- Clarke, C. J. S. (1995). The Nonlocality of mind. *Journal of Consciousness Studies*, 2 (3), 231-240.
- Crick, F. (1994). *The Astonishing Hypothesis*. New York, NY: Scribner.
- Dainton, B. (2006). *Stream of Consciousness: Unity and Continuity in Conscious Experience (2nd ed.)*. Abingdon, UK: Routledge.
- Del Cul, A., Dehaene, S., Reyes, P., Bravo, E. & Slachevsky, A. (2009). Causal role of prefrontal cortex in the threshold for access to consciousness. *Brain*, 132 (Pt 9), 2531-2540. [10.1093/brain/awp111](https://doi.org/10.1093/brain/awp111)
- Doesburg, S. M., Green, J. J., McDonald, J. J. & Ward, L. M. (2009). Rhythms of consciousness: binocular rivalry reveals large-scale oscillatory network dynamics mediating visual perception. *PloS One*, 4 (7), e6142-e6142. [10.1371/journal.pone.0006142](https://doi.org/10.1371/journal.pone.0006142)
- (2012). Theta modulation of inter-regional gamma synchronization during auditory attention control. *Brain Research*, 1431, 77-85. [10.1016/j.brainres.2011.11.005](https://doi.org/10.1016/j.brainres.2011.11.005)
- Engel, A. K., Fries, P., König, P., Brecht, M. & Singer, W. (1999). Temporal binding, binocular rivalry, and consciousness. *Consciousness and Cognition*, 8, 128-151. [10.1006/ccog.1999.0389](https://doi.org/10.1006/ccog.1999.0389)
- Fries, P. (2009). Neuronal gamma-band synchronization as a fundamental process in cortical computation. *Annual Review of Neuroscience*, 32, 209-224. [10.1146/annurev.neuro.051508.135603](https://doi.org/10.1146/annurev.neuro.051508.135603)
- Goupil, L. & Bekinschtein, T. A. (2011). Cognitive processing during the transition to sleep. *Archives Italiennes de Biologie*, 150 (2-3), 140-154.
- Gross, J., Schmitz, F., Schnitzler, I., Kessler, K., Shapiro, K., Hommel, B. & Schnitzler, A. (2004). Modulation of long-range neural synchrony reflects temporal limitations of visual attention in humans. *Proceedings of the National Academy of Sciences*, 101 (35), 13050-13050. [10.1073/pnas.0404944101](https://doi.org/10.1073/pnas.0404944101)
- Gruber, T., Müller, M. M. & Keil, A. (2002). Modulation of induced gamma band responses in a perceptual learning task in the human EEG. *Journal of Cognitive Neuroscience*, 14 (5), 732-744. [10.1162/08989290260138636](https://doi.org/10.1162/08989290260138636)
- Hobson, J. A., Pace-Schott, E. F. & Stickgold, R. (2000). Dreaming and the brain: toward a cognitive neuroscience of conscious states. *Behavioral and Brain Sciences*, 23 (6), 904-1121. [10.1017/S0140525X00003976](https://doi.org/10.1017/S0140525X00003976)
- James, W. (1890). *The Principles of Psychology*. New York, NY: Holt.
- Jouvet, M. (1979). What does a cat dream about? *Trends in Neurosciences*, 2, 280-282. [10.1016/0166-2236\(79\)90110-3](https://doi.org/10.1016/0166-2236(79)90110-3)
- Kelly, E. R. (1882). *The Alternative: A Study in Psychology*. London, UK: Macmillan.
- Kiverstein, J. (2010). Making sense of phenomenal unity: An intentionalist account of temporal experience. *Royal Institute of Philosophy Supplement*, 67, [10.1017/S1358246110000081](https://doi.org/10.1017/S1358246110000081)
- Kornmeier, J. & Bach, M. (2012). Ambiguous figures—what happens in the brain when perception changes but not the stimulus. *Frontiers in Human Neuroscience*, 6, 51-51. [10.3389/fnhum.2012.00051](https://doi.org/10.3389/fnhum.2012.00051)

- Kottlow, M., Jann, K., Dierks, T. & Koenig, T. (2012). Increased phase synchronization during continuous face integration measured simultaneously with EEG and fMRI. *Clinical Neurophysiology*, 123, 1536-1548. [10.1016/j.clinph.2011.12.019](https://doi.org/10.1016/j.clinph.2011.12.019)
- Martini, N., Menicucci, D., Sebastiani, L., Bedini, R., Pingitore, A., Vanello, N., Milanesi, M., Landini, L. & Gemignani, A. (2012). The dynamics of EEG gamma responses to unpleasant visual stimuli: From local activity to functional connectivity. *NeuroImage*, 60 (2), 922-932. [10.1016/j.neuroimage.2012.01.060](https://doi.org/10.1016/j.neuroimage.2012.01.060)
- Massimini, M., Ferrarelli, F., Murphy, M. J., Huber, R., Riedner, B. A., Casarotto, S. & Tononi, G. (2010). Cortical reactivity and effective connectivity during REM sleep in humans. *Cognitive Neuroscience*, 1, 176-183. [10.1080/17588921003731578](https://doi.org/10.1080/17588921003731578)
- Melloni, L., Molina, C., Pena, M., Torres, D., Singer, W. & Rodriguez, E. (2007). Synchronization of neural activity across cortical areas correlates with conscious perception. *Journal of Neuroscience*, 27, 2858-2865. [10.1523/JNEUROSCI.4623-06.2007](https://doi.org/10.1523/JNEUROSCI.4623-06.2007)
- Merker, B. (2007). Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. *Behavioral and Brain Sciences*, 30, 63-81. [10.1017/S0140525X07000891](https://doi.org/10.1017/S0140525X07000891)
- Móró, L. (2010). Hallucinatory altered states of consciousness. *Phenomenology and the Cognitive Sciences*, 9, 241-252. [10.1007/s11097-010-9162-2](https://doi.org/10.1007/s11097-010-9162-2)
- Noreika, V. (2014). *Alterations in the States and Contents of Consciousness: Empirical and Theoretical Aspects*. Turku, Finland: University of Turku.
- Noreika, V., Valli, K., Lahtela, H. & Revonsuo, A. (2009). Early-night serial awakenings as a new paradigm for studies on NREM dreaming. *International Journal of Psychophysiology*, 74, 14-18. [10.1016/j.ijpsycho.2009.06.002](https://doi.org/10.1016/j.ijpsycho.2009.06.002)
- Noreika, V., Jylhäkangas, L., Móró, L., Valli, K., Kaskinoro, K., Aantaa, R., Scheinin, H. & Revonsuo, A. (2011). Consciousness lost and found: Subjective experiences in an unresponsive state. *Brain and Cognition*, 77 (3), 327-334. [10.1016/j.bandc.2011.09.002](https://doi.org/10.1016/j.bandc.2011.09.002)
- Noreika, V., Canales-Johnson, A., Koh, J., Taylor, M., Massey, I. & Bekinschtein, T. A. (Under Review). Intrusions of a drowsy mind: Electroencephalographic correlates of phenomenological unpredictability.
- Nummenmaa, L., Glerean, E., Hari, R. & Hietanen, J. K. (2014). Bodily maps of emotions. *Proceedings of the National Academy of Sciences*, 111, 646-651. [10.1073/pnas.1321664111](https://doi.org/10.1073/pnas.1321664111)
- Osipova, D., Takashima, A., Oostenveld, R., Fernández, G., Maris, E. & Jensen, O. (2006). Theta and gamma oscillations predict encoding and retrieval of declarative memory. *Journal of Neuroscience*, 26 (28), 7523-7531. [10.1523/JNEUROSCI.1948-06.2006](https://doi.org/10.1523/JNEUROSCI.1948-06.2006)
- Parvizi, J. & Damasio, A. (2001). Consciousness and the brainstem. *Cognition*, 79 (1-2), 135-160. [10.1016/S0010-0277\(00\)00127-X](https://doi.org/10.1016/S0010-0277(00)00127-X)
- Pfurtscheller, G., Gramann, B., Huggins, J. E., Levine, S. P. & Schuh, L. A. (2003). Spatiotemporal patterns of beta desynchronization and gamma synchronization in corticographic data during self-paced movement. *Clinical Neurophysiology*, 114 (7), 1226-1236. [10.1016/S1388-2457\(03\)00067-1](https://doi.org/10.1016/S1388-2457(03)00067-1)
- Pöppel, E. (1997). A hierarchical model of temporal perception. *Trends in Cognitive Sciences*, 1 (2), 56-61. [10.1016/S1364-6613\(97\)01008-5](https://doi.org/10.1016/S1364-6613(97)01008-5)
- Railo, H., Koivisto, M. & Revonsuo, A. (2011). Tracking the processes behind conscious perception: A review of event-related potential correlates of visual consciousness. *Consciousness and Cognition*, 20 (3), 972-983. [10.1016/j.concog.2011.03.019](https://doi.org/10.1016/j.concog.2011.03.019)
- Revonsuo, A. (2006). *Inner presence. Consciousness as a Biological Phenomenon*. Cambridge, MA: MIT Press.
- Revonsuo, A., Kallio, S. & Sikka, P. (2009). What is an altered state of consciousness? *Philosophical Psychology*, 22 (2), 187-204. [10.1080/09515080902802850](https://doi.org/10.1080/09515080902802850)
- Rosenthal, D. M. (1986). Two concepts of consciousness. *Philosophical Studies*, 49 (3), 329-359. [10.1007/BF00355521](https://doi.org/10.1007/BF00355521)
- Siclari, F., LaRocque, J. J., Postle, B. R. & Tononi, G. (2013). Assessing sleep consciousness within subjects using a serial awakening paradigm. *Frontiers in Psychology*, 4 (542), 542-542. [10.3389/fpsyg.2013.00542](https://doi.org/10.3389/fpsyg.2013.00542)
- Singer, W. (2001). Consciousness and the binding problem. *Annals of the New York Academy of Sciences*, 929, 123-146. [10.1111/j.1749-6632.2001.tb05712.x](https://doi.org/10.1111/j.1749-6632.2001.tb05712.x)
- (2015). The Ongoing Search for the Neuronal Correlate of Consciousness. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1-30). Frankfurt a. M., GER: MIND Group.
- Sitt, J. D., King, J.-R., Karoui, I. E., Rohaut, B., Faugeras, F., Gramfort, A., Cohen, L., Sigman, M., Dehaene, S. & Naccache, L. (2014). Large scale screening of neural signatures of consciousness in patients in a vegetative or minimally conscious state. *Brain*, 137 (Pt 8), 2258-2270. [10.1093/brain/awu141](https://doi.org/10.1093/brain/awu141)
- Steriade, M., Contreras, D., Amzica, F. & Timofeev, I. (1996). Synchronization of fast (30-40 Hz) spontaneous

- oscillations in intrathalamic and thalamocortical networks. *Journal of Neuroscience*, 16 (8), 2788-2808.
- Tart, C. T. (1972). States of consciousness and state-specific sciences. *Science*, 176, 1203-1210.
[10.1126/science.176.4040.1203](https://doi.org/10.1126/science.176.4040.1203)
- Thompson, E. (2015). Dreamless sleep, the embodied mind, and consciousness: The relevance of a classical Indian debate to cognitive science. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1-20). Frankfurt a. M., GER: MIND Group.
- Tononi, G. (2012). The integrated information theory of consciousness: an updated account. *Archives Italiennes de Biologie*, 150 (2-3), 56-90.
- Treisman, A. (1996). The binding problem. *Current Opinion in Neurobiology*, 6 (2), 171-178.
[10.1016/S0959-4388\(96\)80070-5](https://doi.org/10.1016/S0959-4388(96)80070-5)
- Vaitl, D., Birbaumer, N., Gruzelier, J., Jamieson, G. A., Kotchoubey, B., Kübler, A., Lehmann, D., Miltner, W. H.R., Ott, U., Pütz, P., Sammer, G., Strauch, I., Strehl, U., Wackermann, J. & Weiss, T. (2005). Psychobiology of altered states of consciousness. *Psychological Bulletin*, 131 (1), 98-127.
[10.1037/0033-2909.131.1.98](https://doi.org/10.1037/0033-2909.131.1.98)
- Zhang, L., Gan, J. Q. & Wang, H. (2014). Optimized gamma synchronization enhances functional binding of fronto-parietal cortices in mathematically gifted adolescents during deductive reasoning. *Frontiers in Human Neuroscience*, 8 (430), eCollection 2014-eCollection 2014. [10.3389/fnhum.2014.00430](https://doi.org/10.3389/fnhum.2014.00430)

State or content of consciousness?

A Reply to Valdas Noreika

Wolf Singer

An attempt is made to distinguish between brain states required to support consciousness and the neuronal underpinnings of conscious versus non-conscious processing in an awake, attentive brain, respectively. It is argued that brain states supporting consciousness are characterised by high dimensional dynamics exhibiting a high degree of complexity, implying that conscious states are graded. Different mechanisms determine whether signals are processed at the conscious or sub-conscious level. Thus, there is no unique neuronal correlate of consciousness.

Keywords

Brain dynamics | Complexity | Conscious processing | Conscious state | Content of consciousness | Dimensionality

Author

Wolf Singer

w.singer@brain.mpg.de

Max Planck Institute for Brain
Research (MPI)
Frankfurt a. M., Germany

Commentator

Valdas Noreika

valdas.noreika@mrc-cbu.cam.ac.uk

Medical Research Council
Cambridge, United Kingdom

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

My sincere thanks go to Valdas Noreika for having identified with succinct clarity the weaknesses in our current attempt to identify the neuronal correlates of consciousness (NCC). I would have sincerely appreciated these comments before finalising my manuscript, as they would have forced me to distinguish more clearly between the neuronal underpinnings of the conscious state and the neuronal correlates of conscious versus unconscious processing.

Noreika is absolutely right in pointing out that the search for the mechanisms permitting access to conscious processing falls short of

identifying the NCC proper and, likewise, that the determination of variables required for the maintenance of a conscious state is insufficient if pursued without considering the contents of conscious processing. The mere fact that one can distinguish between the “conscious state” and the conditions required for “conscious processing”, yet also consider both as targets in the search for the NCC, suggests that the explanandum is ill-defined. Presently, both studies devoted to the distinction between conscious and unconscious processing and those investigating the brain states required for conscious pro-

cessing are considered as investigations of the NCC, although they clearly target different neuronal mechanisms. Thus, studies on consciousness are fraught with the problem of a lack of a clear definition of “the” consciousness for which we wish to find a neuronal correlate. Another problem is that we are still far from fully understanding the neuronal mechanisms underlying higher cognitive functions. Behavioural studies suggest, for example, that perception involves probabilistic Bayesian-matching operations in which sensory evidence is compared with stored knowledge about the probability of occurrence and the features of the respective perceptual objects. However, it is entirely unknown where and how the huge amount of priors are stored, how the specific priors can be retrieved on the fly within the few hundreds of milliseconds sufficient for recognition, and how the matching operations are realized in neuronal networks. Thus, at the present stage it is even impossible to precisely define the signatures of neuronal activity that could be considered the result of a perceptual process or as the neuronal representation of a percept.

In the light of these uncertainties, the distinctions between conscious and unconscious processing or between states compatible with conscious and unconscious processing, respectively, appear to be exploited primarily in order to learn more about mechanisms underlying pattern recognition, decision making, and intentionality, rather than serving the search for the neuronal underpinnings of the ill-defined phenomenon that we address as “consciousness”. In contrast to NCC research, these more humble approaches have been quite successful, probably because the explananda are well-defined and can be operationalised.

2 The conscious state

The analysis of the neuronal prerequisites required for the maintenance of consciousness has a long history and has only recently been considered part of consciousness research. The reason for this is that the criteria used for distinctions between conscious and non-conscious states or altered states of consciousness can be

tested in both humans and animals. Examples of these criteria are reactivity to noxious stimuli, the ability to move intentionally, and the ability to accomplish a number of well-defined cognitive tasks, involving attention, short and long term memory, recognition, and decision making. Thus, the plethora of studies performed both on animals and humans on the neuronal mechanisms underlying arousal, attention, wakefulness, sleep, anaesthesia, and coma all contribute to our understanding of the neuronal prerequisites of states permitting conscious processing. Accordingly, it is well-established that brain functions characteristic of the conscious state require that neuronal networks operate in a critical dynamical range. This range is regulated by half a dozen globally-acting modulatory systems that originate in deep and evolutionary ancient brain structures. The adjustable neuronal parameters are essentially the balance between excitatory and inhibitory effects and the time- and length constants of dendritic integration. These adjustments lead to marked modifications of the system’s dynamics. These modulations are reflected by changes in the prevailing frequencies of oscillatory activity, the degree and spatial granularity of synchronisation (also addressed as correlation length), and the propagation of signals across the network.

Classical brain theories have not attributed much attention to the significance of these dynamic variables for processing and assume that loss of consciousness in sleep and anaesthesia is essentially due to reduced excitability and signal transmission. However, in more recent theories, brain dynamics are thought to play a crucial role in information processing. This novel framework provides much more specific explanation of the breakdown of consciousness in sleep, anaesthesia, and coma. These theories posit that oscillations and the concomitant variables, such as synchronisation, phase locking, phase relations, and cross frequency coupling, are relevant for signal selection by attention, binding operations, and the representation of nested semantic relations (for review see [Singer 1999](#); [Buzsáki et al. 2013](#)). In addition, these complex dynamics have been proposed as a substrate for the generation of the high-di-

mensional coding space required for the storage and superposition of priors, the matching of stored information with sensory evidence, and the segregation of patterns for classification (for review see [Singer 2013](#)). The basis of these operations is the transformation of low-dimensional input patterns into high-dimensional dynamic states, in order to perform the necessary computations in this space and to then retransform the results into low-dimensional output signals. The advantages of performing computations in high-dimensional dynamic space are currently explored in the conceptual framework of “reservoir computing” or “liquid state or echo state machines” ([Bertschinger & Natschläger 2004](#); [Buonomano & Maass 2009](#); [Jaeger 2001](#)).

Recent analysis of the properties of recurrent networks, such as those realized in neuronal systems and in particular the cerebral cortex, indicate that such high-dimensional dynamic states can indeed be generated in delay-coupled networks ([Lazar et al. 2009](#); [Buonomano & Maass 2009](#); [Soriano et al. 2013](#); for review see [Singer 2013](#)). In the present context it is important to recall that the dynamics required for such computations can emerge only when the networks are in the appropriate state. The optimal state has been identified as the edge of chaos, slightly below self-organised criticality, the so-called SOC state, because in this state the dimensionality or the complexity of the system are very high. Computationally this range is optimal because it offers a maximum of possible bifurcation points and storage capacity. ([Plenz & Thiagarajan 2007](#)). In this conceptual framework, computational results should consist of substates with reduced dimensionality. Experimental evidence indicates that the high-dimensional resting states are actually reduced by sensory input, imagery, recall of memories, or focused attention. These processes are all associated with enhanced correlation between neuronal responses due to the induction of synchronized high-frequency oscillations—where enhancing correlations reduces dimensionality (for review see [Singer 2013](#)). The notion that SOC states are optimal prerequisites for processing also fits with the robust evidence that states compatible with consciousness are characterized

by “desynchronized” brain activity, i.e., states characterized by uncorrelated activity, such as are typical for wakefulness and arousal. If, and evidence suggests this to be the case (for review see [Singer 1999, 2013](#)), establishment of lower-dimensional synchronous substates, e.g., the formation of transiently-synchronized assemblies of neurons, is an integral part of the computations, then dynamic states characterized by global, large scale synchrony would be inappropriate as background for computations underlying higher cognitive functions.

As outlined in the target paper and above, higher cognitive functions require fine-grained binding operations among semantically-related contents that need to be encoded in ad hoc-formed neuronal assemblies. Such concatenation of multiple assemblies by partial correlations and perhaps also cross-frequency coupling would be impossible in networks that are already highly synchronized to begin with and hence exhibit low complexity and dimensionality. The well-established notion that deep sleep, anaesthesia, and most forms of coma are associated with brain states that exhibit slow oscillations synchronized over considerable distances agrees with this interpretation. In agreement with the prediction that low-dimensional brain states are incompatible with sophisticated processing are also the recent stimulation experiments cited by Noreika. It is to be expected that stimulation of a dynamic system that is in a low-dimensional state and at an overall reduced level of excitability will elicit only a spatially-restricted responses of low complexity—in particular if the stimulus is itself very low-dimensional, as is the case for a TMS pulse.

Considering more recent theories on brain functions, it appears as if the prerequisite or the NCC of a conscious state is a dynamic state that assures a high degree of complexity and high-dimensionality of resting-state dynamics. It is only in this state that the higher cognitive functions can be realized that one expects from a conscious brain.

It should be noted, however, that this operational definition of consciousness makes no inferences about the subjective contents of consciousness or the awareness of particular qualia

of experience. According to this definition, consciousness is simply a brain state that allows animals and humans to accomplish higher cognitive functions that include not only perception but also decision making, planning of actions, generation of procedural and episodic memories, and last but not least intentionality and reasoning. Thus, one would expect consciousness, defined in this way, to be a graded phenomenon. If the state of the brain changes towards reduced complexity and dimensionality, there should be a graded deterioration of functions. Those requiring integration of widely-distributed assemblies should become impeded first, while simple reactions to salient sensory stimuli would persist for much longer. This seems to be in perfect agreement with the gradual deterioration of cognitive functions as the brain state shifts from high levels of alertness to drowsiness and sleep.

3 Conscious versus subconscious processing

As Noreika points out, “consciousness” defined by the status of phenomenal content is something very different from a conscious state, as this connotation of consciousness can only be investigated in human subjects. The reason for this is that the distinguishing criterion is the degree of subjective awareness of a cognitive content, and this variable can only be assessed through verbal report. It is simply not possible to know whether a monkey trained to press a lever to signal that it has recognized a particular pattern has the subjective experience that we equate with conscious perception. The monkey brain has the same mechanisms as humans for the allocation of attention, the selection of objects for perception, and the routing of experiences to the different storage modes (working memory, procedural and episodic memory). Thus it is very likely that monkeys are aware of their perceptions in a similar way to us, and that the distinction between conscious and non-conscious processing holds for them as well—but we have no way of knowing. Conditioned lever presses in response to stimuli do not require conscious perception of the stimuli, just as

stopping at a red light while being engaged in a conversation does not require conscious recollection of having perceived the light. It is for this reason that the criterion for conscious processing is the reportability of the perceived stimulus, and hence this aspect of consciousness can only be studied in humans.

Attempts to identify the differences between the neuronal processes that accompany non-conscious and conscious processing, respectively, are of course interesting in their own right. The expectation is that they will provide answers to the question of why certain processes are reportable and have access to working and episodic memory while others are excluded, or the question of why certain forms of reasoning and decision-making require conscious deliberations while others do not. However, as pointed out so stringently by Noreika, these attempts fall short of identifying the NCC proper, and at best cover some aspects of conscious processing while being fraught with problems. The most difficult problems are related to the distinction between the processes that are essential for subjective awareness and reportability and those that are the consequence of having become aware of something or that simply provide favourable conditions for becoming aware, such as the allocation of attention or the saliency of stimuli. So far the only neuronal signatures distinguishing between reportable and non-reportable processes have been found to be transitory, lasting at most a few hundred milliseconds. Noreika argues rightly that this disqualifies these events as NCCs because the stream of consciousness is continuous and the awareness of contents can persist for quite some time.

4 Conclusion and outlook

We need to be more cautious when using the term NCC and to define precisely, each time we perform a search for underlying neuronal mechanisms, which of the many aspects of “consciousness” we actually intend to investigate. We need to differentiate between processes assuring access to conscious processing, which are expected to be transient, and processes necessary for sustaining the stream of consciousness

that has longer time-constants. And finally, we need to distinguish processes assuring sustained awareness of contents that are most likely related to the transfer of material to short- and long-term memories. If we proceed in this way, subdividing “consciousness” into subfunctions including reportability and defining these as explananda, some of the present problems may dissolve. However, the consequence is that we shall have to give up the search for “the” overarching NCC.

If we pursue this agenda, it is to be expected that correlates will be found for all aspects of consciousness except those associated with the “hard” problem, which appears to be a specific human problem. As I argued in the target paper, searching for the neuronal correlates of qualia in individual brains is unlikely to be successful because the immaterial and therefore somewhat mysterious connotations of qualia are likely to have the status of social realities. What we can achieve, however, is an identification of brain processes that underlie those cognitive functions required for generating social realities. These would be the ability to engage in social interaction, to develop a theory of mind, to find symbolic descriptions of internal states, and to reach consensus on the “reality” of these through communication with others.

To conclude this brief reply to the extremely inspiring commentary on my target paper, I want to express my sincere gratitude to Noreika for having pointed out the critical issues in our research on the NCC. The reply forced me to engage with this research again and helped me substantially in clarifying my own position in the debate.

References

- Bertschinger, N. & Natschläger, T. (2004). Real-time computation at the edge of chaos in recurrent neural networks. *Neural Computation*, 16 (7), 1413-1436.
[10.1162/089976604323057443](https://doi.org/10.1162/089976604323057443)
- Buonomano, D. V. & Maass, W. (2009). State-dependent computations: Spatiotemporal processing in cortical networks. *Nature Reviews Neuroscience*, 10, 113-125.
[10.1038/nrn2558](https://doi.org/10.1038/nrn2558)
- Buzsáki, G., Logothetis, N. & Singer, W. (2013). Scaling brain size, keeping timing: Evolutionary preservation of brain rhythms. *Neuron*, 80 (3), 751-764.
[10.1016/j.neuron.2013.10.002](https://doi.org/10.1016/j.neuron.2013.10.002)
- Jaeger, H. (2001). The “echo state” approach to analysing and training recurrent neural networks - with an Erratum note. *German National Research Center for Information Technology, GMD Report*, 148
- Lazar, A., Pipa, G. & Triesch, J. (2009). SORN: A self-organizing recurrent neural network. *Frontiers in Computational Neuroscience*, 3 (23), 1-9.
[10.3389/neuro.10.023.2009](https://doi.org/10.3389/neuro.10.023.2009)
- Plenz, D. & Thiagarajan, T. C. (2007). The organizing principles of neuronal avalanches: Cell assemblies in the cortex? *Trends in Neurosciences*, 30 (3), 99-110.
[10.1016/j.tins.2007.01.005](https://doi.org/10.1016/j.tins.2007.01.005)
- Singer, W. (1999). Neuronal synchrony: A versatile code for the definition of relations? *Neuron*, 24 (1), 49-65.
[10.1016/S0896-6273\(00\)80821-1](https://doi.org/10.1016/S0896-6273(00)80821-1)
- (2013). Cortical dynamics revisited. *Trends in Cognitive Sciences*, 17 (12), 616-626.
[10.1016/j.tics.2013.09.006](https://doi.org/10.1016/j.tics.2013.09.006)
- Soriano, M. C., Garcia-Ojalvo, J., Mirasso, C. R. & Fischer, I. (2013). Complex photonics: Dynamics and applications of delay-coupled semiconductor lasers. *Review of Modern Physics*, 85 (1), 421-470.
[10.1103/RevModPhys.85.421](https://doi.org/10.1103/RevModPhys.85.421)

Dreamless Sleep, the Embodied Mind, and Consciousness

The Relevance of a Classical Indian Debate to Cognitive Science

Evan Thompson

One of the major debates in classical Indian philosophy concerned whether consciousness is present or absent in dreamless sleep. The philosophical schools of Advaita Vedānta and Yoga maintained that consciousness is present in dreamless sleep, whereas the Nyāya school maintained that it is absent. Consideration of this debate, especially the reasoning used by Advaita Vedānta to rebut the Nyāya view, calls into question the standard neuroscientific way of operationally defining consciousness as “that which disappears in dreamless sleep and reappears when we wake up or dream.” The Indian debate also offers new resources for contemporary philosophy of mind. At the same time, findings from cognitive neuroscience have important implications for Indian debates about cognition during sleep, as well as for Indian and Western philosophical discussions of the self and its relationship to the body. Finally, considerations about sleep drawn from the Indian materials suggest that we need a more refined taxonomy of sleep states than that which sleep science currently employs, and that contemplative methods of mind training are relevant for advancing the neurophenomenology of sleep and consciousness.

Keywords

Access consciousness | Advaita vedānta | Anaesthesia | Awareness | Buddhism | Consciousness | Cross-cultural philosophy of mind | Dreamless sleep | Meditation | Memory | Neurophenomenology | Nrem (non-rapid eye movement) sleep | Nyāya | Phenomenal consciousness | Self | Self-experience | Sleep | Yoga

1 Introduction

Many neuroscientists and philosophers today think of dreamless sleep (see [glossary](#)) as a blackout state in which consciousness is entirely absent. Indeed, they often appeal to this apparent fact in order to define consciousness:

Everybody knows what consciousness is: it is what vanishes every night when we fall into a dreamless sleep and reappears when we wake up or when we dream. (Tononi 2008, p. 216)

Consciousness consists of inner, qualitative, subjective states and processes of sentience and awareness. Consciousness, so defined, begins when we wake in the morning from a dreamless sleep and continues until we fall asleep again, die, go into a coma, or otherwise become “unconscious”. (Searle 2000, p. 559)

I will call the view that consciousness vanishes or ceases in dreamless sleep the *default view* of the relationship between consciousness and dreamless sleep. One aim of this paper is to argue that the

Author

[Evan Thompson](#)

evan.thompson@ubc.ca

University of British Columbia
Vancouver, BC, Canada

Commentator

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

Glossary

1. Canonical physiological sleep states according to polysomnography	<p><i>“Light Sleep”</i></p> <ul style="list-style-type: none"> • Stage 1: closed eyes, slow eye-rolling movements, EEG alpha waves (8–12 Hz) subside, slower theta waves (4–8 Hz) arrive. • Stage 2: eye movements cease, 12–14 Hz bursts (sleep spindles) and brief high voltage waves (K-complexes) occur. <p><i>“Deep Sleep” or “Slow-Wave Sleep”</i></p> <ul style="list-style-type: none"> • Stage 3: a mixture of sleep spindles and high-amplitude, slow frequency delta waves (0.5–4 Hz). • Stage 4: delta waves almost exclusively. • REM (Rapid Eye Movement) or “Paradoxical Sleep”: fast-frequency, low-amplitude waves, limb muscles paralyzed, eyes closed with rapid eye movements.
2. Phenomenological sleep terms	<ul style="list-style-type: none"> • Sleep mentation: sleep thoughts and images. • Dreaming: immersion in the imagined dreamworld; “immersive spatiotemporal hallucination” (Windt 2010). • Lucid Dreaming: knowing that one is dreaming while dreaming; being able to direct one’s attention to the dream as a dream (Windt & Metzinger 2007). • Dreamless sleep (Western conception): sleep lacking mentation. • Dreamless sleep (Indian conception): sleep lacking mentation; phenomenal character of peaceful, non-intentional awareness. • Lucid dreamless sleep (Indian conception): sleep lacking mentation; phenomenal character of peaceful, non-intentional awareness; non-conceptual meta-awareness (“witness consciousness”) of the dreamless sleep state.

Glossary of Indian philosophical systems

CONSCIOUSNESS IN DREAMLESS SLEEP	
Yoga	<ul style="list-style-type: none"> • <i>Yoga Sūtras</i>, traditionally ascribed to Patañjali, though authorship is uncertain (c. 3rd–4th century CE). The commentary attributed to Vyāsa may in fact have been written by Patañjali.
Advaita Vedānta (Advaitins)	<ul style="list-style-type: none"> • Śaṅkara (788–820 CE). • Sureśvara (c. 9th century CE). • Madhusūdana (c. 16th century CE).
Buddhism	<ul style="list-style-type: none"> • The Theravāda school postulates a basal and passive “life continuum” or “factor of existence” consciousness (<i>bhavaṅga</i>) that occurs in dreamless sleep (c. 3rd century BCE–2nd century CE). • The Yogācāra school postulates a basal “store consciousness” (<i>ālaya-vijñāna</i>), which persists in dreamless sleep (c. 4th century CE).
NO CONSCIOUSNESS IN DREAMLESS SLEEP	
Nyāya (Nyaiyāyikas)	<ul style="list-style-type: none"> • <i>Nyāya Sūtras</i>, authored by Gautama (c. 2nd century BCE). • Vātsyāyana (c. 450 CE). • Udyotakara (c. 550 CE). • Udayana (c. 10th century CE).

default view is not as obvious or strong as it is often thought to be. Another aim is to propose that we need a finer taxonomy of sleep states than that which sleep science currently employs, in order to allow for the possibility of states or phases of dreamless sleep in which consciousness is present. There are forceful reasons, if not decisive ones, for describing certain kinds of dreamless sleep as modes of consciousness rather than as the absence of consciousness. These reasons derive from the debate about dreamless sleep between the Advaita Vedānta and Nyāya schools of Indian philosophy (see [glossary](#)). Examining this debate in the light of cognitive science raises important conceptual and methodological issues for the cognitive neuroscience of consciousness. Furthermore, considerations about sleep drawn from Indian philosophy suggest new experimental questions and protocols for the cognitive neuroscience of sleep and consciousness. By weaving together these different traditions—Western cognitive science and Indian philosophy—I hope to show the value of cross-cultural philosophy of mind for cognitive science.

2 The experience of waking up

Before turning to the Indian debate, I would like to motivate the examination of dreamless sleep and consciousness by considering the experience of waking up from deep sleep and what this experience reveals about our experience of the self.

One of the best descriptions of waking up comes from Marcel Proust. In a long passage at the beginning of the first volume of *In Search of Lost Time*, the unnamed narrator describes awakening from sleep:

A sleeping man holds in a circle around him the sequence of the hours, the order of the years and world. He consults them instinctively as he wakes and reads in them in a second the point on the earth he occupies, the time that has elapsed up to his waking; but their ranks can be mixed up, broken. If towards morning, after a bout of insomnia, sleep overcomes him as he is reading, in a position too different from

the one in which he usually sleeps, his raised arm alone is enough to stop the sun and make it retreat, and, in the first minute of his waking, he will no longer know what time it is, he will think he has only just gone to bed. If he dozes off in a position still more displaced and divergent, for instance after dinner sitting in an armchair, then the confusion among the disordered worlds will be complete, the magic armchair will send him travelling at top speed through time and space, and, at the moment of opening his eyelids, he will believe he went to bed several months earlier in another country. But it was enough if, in my own bed, my sleep was deep and allowed my mind to relax entirely; then it would let go of the map of the place where I had fallen asleep and, when I woke in the middle of the night, since I did not know where I was, I did not even understand in the first moment who I was; all I had, in its original simplicity, was the sense of existence as it may quiver in the depths of an animal; I was more bereft than a cave-man; but then the memory—not yet of the place where I was, but of several of those where I had lived and where I might have been—would come to me like help from on high to pull me out of the void from which I could not have got out on my own; I passed over centuries of civilization in one second, and the image confusedly glimpsed of oil lamps, then of wing-collar shirts, gradually recomposed my self's original features. ([Proust 2003](#), p. 9)

Proust depicts the moment of awakening from deep sleep as one where we have lost all sense of the self derived from memories of the episodes of our lives. Instead of the autobiographical or narrative sense of self as a person with a storyline through time, there remains only the sensation of existing at that moment. What marks the first instant of awakening is not the self of memory but the feeling of being alive, or what Proust calls “the sense of existence as it may quiver in the depths of an animal.”

The moment of awakening thus reveals two kinds of self-experience. The first kind is the embodied self-experience of being alive in the present moment, or the experience of being sentient. The second kind of self-experience is the autobiographical experience of being a person with a storyline, a thinking being who mentally travels in time. The first kind of embodied sense of self we experience immediately upon awakening, but as we reach automatically for the second kind of autobiographical sense of self, it sometimes goes missing.

This distinction between two modes of self-experience, one of which remains present in the sleep–wake transition even if the other is lost, suggests the following tentative phenomenological line of thought leading towards the idea of consciousness being present in certain phases of dreamless sleep.

Consider that although deep sleep creates a gap or a rupture in our consciousness, we often feel the gap immediately upon awakening. Our waking sense that we were just asleep and unknowing is not outside knowledge—like the kind we have when we know about someone else’s having been asleep; it is inside, first-hand experience. We are aware of the gap in our consciousness from within our consciousness. Although we may forget many things about ourselves when we first wake up—where we are, how we got there, maybe even our name—we do not have to turn around to see who it was who was just asleep and unknowing, if by “who” we mean the sense of self as the embodied subject of present-moment experience in contrast to the sense of self as the mentally represented object of autobiographical memory. This intimate and immediate bodily self-awareness that we have as we emerge from sleep into waking life suggests that there may be some kind of deep-sleep awareness, operative at least for some stretch of time prior to waking up, a taste of which we retain in the waking state, despite there being no specific mental content to recall. If there is a deep-sleep awareness we can retain in this way, then there may, at least for certain phases of deep sleep, be a phenomenal character to deep sleep or something “it is like” (Nagel 1974) to be deeply asleep—in which case consciousness

cannot be entirely absent from deep sleep (Sharma 2001).

This line of thought finds its strongest philosophical expression in classical Indian philosophy, so if we wish to see whether we can sharpen it into a more compelling argument, we need to look at the Indian discussions.

3 A classical Indian debate

In the earliest texts of the *Upaniṣads*, dating from the seventh century B.C.E., dreamless sleep is singled out as one of the principal states of the self, along with the waking state and the dream state. Various characterizations of dreamless sleep are given. Some texts characterize it as a state of oblivion, while other texts describe it as a mode of unknowing or non-cognitive consciousness that lacks either the outer sensory objects of the waking state or the inner mental images of the dream state (Raveh 2008). It is this second characterization that we find in the later texts of the Yoga and Vedānta schools. These texts also present a basic form of philosophical argument for dreamless sleep being a mode of consciousness. The argument runs as follows: When you wake up from a dreamless sleep, you are aware of having had a peaceful sleep. You know this directly from memory, so the argument asserts, not from inference. In other words, you do not need to reason, “I feel well rested now, so I must have had a peaceful sleep.” Rather, you are immediately aware of having been happily asleep. Memory, however, presupposes the existence of traces that are themselves caused by previous experiences, so in remembering that you slept peacefully, a peaceful feeling must have been experienced. To put the thought another way, the memory report, “I slept peacefully,” would not be possible if awareness were altogether absent from deep sleep; but to say that awareness is present in deep sleep is to say that deep sleep is a mode of consciousness.

To my knowledge, the earliest version of this argument comes from Vyāsa’s third or fourth century C.E. commentary on Patañ-

jali's *Yoga Sūtras*.¹ Patañjali defines yoga as the stilling or restraining of the “fluctuations” of consciousness (*Yoga Sūtras* I:2). When this stilling is accomplished, the “seer” or “witness” can abide in its true form, namely, pure awareness; otherwise the “seer” identifies with the fluctuations of consciousness—with the movements of thought and emotion (I:3–4). Patañjali identifies five kinds of fluctuations or changing states of consciousness: correct cognition, error, imagining or conceptual construction, sleep, and memory (I:5–6), and he defines sleep as a state of consciousness that is based on an “absence” (I:10).

As the traditional commentaries indicate, “absence” does not mean absence of consciousness; it means absence of an object presented to consciousness.² Deep and dreamless sleep is a kind of consciousness without an object. When we are awake we cognize outer objects, and when we dream we cognize mental images. When we are deeply asleep, however, we do not cognize anything—there is no object being cognized and no awareness of oneself as knower. Nevertheless, according to Yoga, we feel this peculiar absence while we sleep and we remember it upon awakening, as evidenced by our saying, “I slept peacefully and I did not know anything.”

Before we examine the debate arising from this argument, let me mention an obvious objection that would occur to us today, especially given what we know from sleep science. The objection is that retrospective subjective evaluations of sleep may be unreliable (Baker et al. 1999), so we cannot assume that the subjective feeling upon awakening of having slept peacefully is based on a veridical memory of a peaceful sleep. An extreme case of the unreliability of self-reports about sleep comes from insomnia patients (Perlis et al. 1997; Rosa & Bonnet 2000; Zhang & Zhao 2007). These patients frequently display sleep-state misperception; that is, their subjective

assessments of the quantity and quality of their sleep deviate strongly from the objective, polysomnographic measures. For example, they often identify themselves as having been awake when they are woken up from polysomnographically-defined sleep, they tend to overestimate sleep-onset latency (the length of time it takes to go from full wakefulness to sleep), and to underestimate total sleep time as compared with polysomnographic measures (Perlis et al. 1997). Even in healthy individuals, the feeling of having slept well could sometimes deviate from objective measures. One could feel refreshed upon awakening, yet the objective measures might show that one's sleep was physiologically restless or intermittent; or one could feel fatigued upon awakening, yet the objective measures might show that one's sleep was physiologically deep and undisturbed. In short, although it is conceptually true that a veridical episodic memory implies having undergone an experience whose content corresponds, to some degree, to that of the memory, it is an empirical matter whether or to what degree any given waking memory impression of sleep is veridical. It is also an empirical question whether episodes of peaceful sleep typically lead to the awareness of having slept peacefully and whether this feeling can occur even when sleep is disturbed.

This line of thought, however, is not decisive against the Yoga argument. Strictly speaking, all this argument needs is the possibility of there being veridical waking memories of having been deeply and dreamlessly asleep in order logically to establish that awareness can be present in at least certain phases or types of dreamless sleep. The argument does not need to establish that waking memory impressions are typically veridical, only that they can be. Indeed, as we will see later, the Yoga viewpoint can allow that ordinary sleep-state perception and retrospective subjective sleep-state evaluations may be unreliable. I will come back to this point at the end of the paper.

A more direct objection to the argument, however, is to challenge the premise that wak-

¹ For a translation of the *Yoga Sūtras* with Vyāsa's commentary, see Āraṇya (1983). Other useful translations can be found in Arya (1989); Bryant (2009); Chapple (2008); Iyengar (1996); and Phillips (2009).

² Arya (1989, pp. 178–184); Bryant (2009, pp. 41–43); Iyengar (1996, pp. 59–60).

ing retrospective reports of sleep are ever memory reports. The philosophers of the Nyāya school (Naiyāyikas) make this challenge. They maintain that the statement, “I slept peacefully and I did not know anything,” expresses an inferential cognition, not a memory report, and that consciousness is entirely absent in dreamless sleep. Given how one feels upon awakening, one infers one had a peaceful sleep and no memory of any dreamless sleep awareness is involved.

Advaita Vedānta, in turn, argues against the Nyāyan viewpoint. The debate between them focuses in particular on the ignorance occurring in dreamless sleep, and specifically on how we know or establish the waking report, “I knew nothing.” While we are asleep we know nothing of this ignorance; we come to know it only upon waking up. Yet given that we do not remain ignorant of our own ignorance, how is this knowing of not-knowing possible? The Naiyāyikas claim that we infer we were ignorant because we do not remember anything, but the Advaitins argue that retrospective oblivion is no proof of a prior lack of consciousness. Moreover, when we wake up we have the feeling of having been asleep and having not known anything. This feeling, the Advaitins claim, is better regarded as a kind of memory brought about by the traces of previous experience. So, in some sense, we must experience our ignorance—the unknowing stillness of our mind—in dreamless sleep.

In reply, the Naiyāyikas claim that we have no consciousness in dreamless sleep, and that when we wake up we make an inference by reasoning in the following way: “While I was in deep sleep, I knew nothing, because I was in a special state (I was not awake) and I lacked the necessary means for knowledge (my senses and mental faculties were shut down).” Of course, the Naiyāyikas are not saying that we explicitly make this inference when we wake up. What they are saying is that what looks like memory is really a case of implicit reasoning taking this inferential form.³

³ My account of the Nyāyan position and of the Advaita Vedānta rebuttal relies heavily on Gupta (1995, pp. 56–66, 99), and Gupta (1998, pp. 84–86). My account simplifies a number of the complexities on both sides of the debate.

In order to understand the kind of inference that the Naiyāyikas think we make, as well as why the Advaitins reject the Nyāyan position, it will be helpful to state the inference in the form of the standard Nyāyan syllogism, which forms an important part of the Nyāyan theory of inferential knowledge.

Suppose we are looking at a hill and you say to me, “There is fire on the hill.” I doubt what you say, however, so you need to convince me. You point to the hill and say, “There is smoke on the hill.” I see the smoke and I am convinced. According to the Nyāya, if we want to unpack how perception and inference have worked together to convince me that you are right, we need to formulate the inferential cognition in the following five steps:

1. There is fire on the hill.
[This is the proposition to be proven. It is what you think when you look at the hill, and it is what you want to convince me is the case.]
2. Because there is smoke on the hill.
[This is the reason you give to support what you say.]
3. Wherever there is smoke there is fire.
[This step states the universal concomitance between the presence of smoke and the presence of fire.]
4. As in the case of the kitchen.
[This step provides an example or actual case of the concomitance, to which we both agree.]
5. There is fire on the hill.
[This step states the conclusion, which is the proposition with which we began, but now stated as established and generated by the preceding inferential process.]

Let us now take this five-step syllogism and apply it to the case of dreamless sleep.⁴ The Nyāya view is that our knowledge that we knew nothing in dreamless sleep is based on the following sort of inference:

1. While I was in dreamless sleep, I knew nothing (there was an absence of knowledge in my self).

⁴ The following inference is my reconstruction of the Naiyāyikas’ reasoning as understood by their Advaita Vedāntin opponents. See Gupta (1995, pp. 56–66, 99), and Gupta (1998, pp. 84–86).

2. This is because (i) I (my self) was in a special state (that is, not awake) or (ii) I (my self) lacked the necessary means for knowledge (that is, my senses and mental faculties were shut down).

3. Whenever (i) I (my self) am in a special state (whenever I am not awake) or (ii) I (my self) lack the necessary means for knowledge (whenever my senses and mental faculties are shut down), I know nothing (there is an absence of knowledge in my self).

4. As in the case of fainting or a blow to the head.

5. While I was in dreamless sleep, I knew nothing (there was an absence of knowledge in my self).

Notice the parallel between the previous inference concerning fire and the present inference concerning dreamless sleep. In the previous case, our concern is to establish the presence of fire on the hill. In the present case, our concern is to establish the absence of knowledge in the self during dreamless sleep. Nevertheless, the form of reasoning is the same.

Again, the Naiyāyikas are not saying that we explicitly go through this inference step by step when we wake up. What they are saying is that we know by inference that we were ignorant during dreamless sleep, and that our inference can be shown to be correct when we make explicit all the steps that it contains. So there is no need to suppose that there is any kind of consciousness during dreamless sleep.

The Advaitins respond by arguing that this inference is faulty and cannot be how we know that there is an absence of knowledge during sleep. The problem is that I need some way to know or establish the reasons for inferring that I knew nothing—namely, that I was in a special state and that I lacked the means for knowledge—and there seems to be no way for me to do this without my relying on the kind of memory these reasons were supposed to obviate.

The first reason the Naiyāyikas give for me to infer that I knew nothing is that I was in a special state, that is, a state different from the waking state. But how do I know that I was

in a special state? If I say, “Because I knew nothing in this state,” then I am reasoning in a circle.

The second reason the Naiyāyikas give for me to infer that I knew nothing is that the means for knowledge were lacking—that is, that my senses and mental faculties were shut down. But here too we need to ask, how do I know that these means were lacking? How do I know my senses and mental faculties were inactive?

Suppose I say, “I infer my senses were shut down because they feel refreshed when I wake up.” But here the same basic problem repeats itself. How do I know or establish that there is a relationship between my senses feeling refreshed and their previously having been inactive? Would I not need to have some experience of knowing that my senses were inoperative together with an experience of knowing I feel refreshed in order to establish a relationship between the two? But while I am asleep I do not have any experience of knowing my senses are inactive; I know this only upon awakening. So how do I establish this relationship? If I appeal to yet another inference, then it looks like I am headed off on an infinite regress.

More generally, the only way I can know that the means for knowledge were absent in deep sleep is by knowing that there was no knowledge present in this state. Only by knowing the effect—my not knowing anything—can I infer the cause—the absence of the means for knowledge. So unless I already know what the inference is trying to establish—that I knew nothing—I cannot establish the reason on which the inference relies.

The Advaita Vedānta conclusion is that I know on the basis of memory, not inference, that I knew nothing in deep sleep. In other words, I remember having not known anything. But a memory is of something previously experienced, so the not-knowing must be experiential.

It is important to highlight the larger metaphysical disputes about the self and cognition that drive this debate. For the Naiyāyikas, the self is a non-physical substance. Unlike Descartes, however, who held that consciousness is the essence of the non-physical mind, the

Naiyāyikas maintain that the self is the substratum of consciousness and that consciousness is an adventitious quality of this substratum that is present only given the appropriate causal conditions, namely when the sensory and mental faculties are functioning to cognize objects. In addition, cognition consists in the taking of a separate object as content and never in taking itself as its own content.⁵ (In the case of introspection, a second-order cognition takes a separate first-order cognition as its object.) For the Advaitins, however, the self is pure consciousness, that is, sheer witnessing awareness distinct from any changing cognitive state. Thus, unlike the Naiyāyikas, the Advaitins cannot allow that consciousness disappears in dreamless sleep, since they think (as do the Naiyāyikas) that it is one and the same self who goes to sleep, wakes up, and remembers having gone to sleep. In addition, for the Advaitins, cognition consists in a reflexive awareness of its own occurrence as an independent prerequisite for the cognition of objects (Ram-Prasad 2007). In other words, the defining feature of cognition is reflexivity or self-luminosity, not intentionality (object-directedness), which is adventitious. Thus, during dreamless sleep, although object-directed cognition is absent, consciousness as reflexive and objectless awareness remains present.

It may help to use the modal notions of necessity and possibility to describe the difference between these views. For the Naiyāyikas, to be in a conscious state is to be in an object-directed state. Given that dreamless sleep is not an object-directed state, it is necessarily the case that consciousness is absent from this state. Nevertheless, if it could be shown that object-directed cognition can occur in dreamless sleep, then the Nyāya could allow for the possibility of consciousness during dreamless sleep. Such consciousness, however, would have to be intermittent or episodic, since object-directed cognitions come and go. What the Nyāya cannot allow is that consciousness is intrinsically reflexive or self-revealing (self-luminous), or that it can occur without an object. Furthermore, for the Nyāya, consciousness requires a

substratum, since consciousness is a mental quality, and mental qualities require the substratum of the self. Therefore, although the self continues to be present during dreamless sleep, consciousness is absent. The Advaitins agree with the Naiyāyikas that the self remains continuously present during dreamless sleep, but they maintain that the self is pure consciousness—consciousness as intrinsically reflexive and self-revealing, not as contingently and adventitiously object-directed. So, for the Advaitins, consciousness cannot ever be absent from dreamless sleep, which is to say that it is necessarily the case that consciousness is present throughout dreamless sleep.

Given these differences, the Nyāya might be thought to be more flexible than Advaita Vedānta with regard to the specific issue about dreamless sleep, since the Nyāya can allow for the possibility of intermittent consciousness during dreamless sleep, whereas Advaita Vedānta cannot allow for any absence of consciousness in this state.

Despite this limitation of the Advaita Vedāntan view, it is possible to extract a key phenomenological idea from its metaphysical commitments. This idea is that when I wake up from a dreamless sleep, it seems that I can sometimes knowingly say I have just emerged from a dreamless sleep, and this saying seems to be a reporting of my awareness, not the product of having to reason things out (Kesarcordi-Watson 1981). It is this thought that provides a premise of the Advaita Vedāntan argument for consciousness continuing in dreamless sleep, and this thought is logically distinct from the Vedāntan belief that the self is essentially pure consciousness.

This phenomenological thought, however, is open to the objection that, given an apparent memory, it does not follow that the state apparently remembered was consciously experienced. For example, we may have apparent memories of childhood events, yet their presence does not imply that these events were consciously experienced, for the memory impressions may have been acquired from other sources of information, such as things our parents told us or family photographs. Similarly, during dreamless

⁵ See Ram-Prasad (2007, Ch. 2) for discussion of the different Indian views about the nature of cognition and consciousness.

sleep, information may accumulate non-consciously from a variety of interoceptive and exteroceptive sources, and upon awakening we may realize that something was going in our mind while we were asleep, though at the time we had no experience of it.

At one level—the level of the empirical psychology of memory—we can make the same reply here that we made above to the objection to the Yoga argument, namely that all the argument requires is the possibility of there being genuine veridical episodic memories upon awakening of having been peacefully asleep; the argument does not need to establish that every apparent waking memory is such a memory. Unlike remote memory (of the sort we have for childhood events) or semantic memory (memory for learned facts or words), episodic memory is standardly taken to require that the events “encoded” in memory are experienced at the time of encoding. So, if there are possible cases upon awakening in which there is any kind of genuine episodic memory “retrieval” of the dreamless-sleep state, it follows that in such cases something about the state of being dreamlessly asleep must have been experientially encoded.

At another level—the level of cross-cultural philosophy of mind—we can see in the Vedāntan phenomenology the basis of a transcendental argument. Transcendental arguments aim to deduce what must be the case in order for some aspect of our experience to be possible. In the present case, the aspect of experience with which we are concerned is not simply that we sleep but that we know that we sleep. What are the necessary conditions of possibility for this kind of self-knowledge? To put the question in a more phenomenological way, how is it possible for you as a conscious subject to experience yourself as one and the same being who falls asleep, who does not actively know anything in being asleep, and who emerges from sleep into waking life? The Vedāntan view is that a retrospective inference across the gap of a complete absence of consciousness will not suffice to make this kind of unified self-experience possible. Rather, you must have some kind of experiential acquaintance with dreamless sleep as a mode of your conscious being.

We can take a further step and think about the Vedāntan argument not just from a Kantian transcendental perspective but also from a Husserlian transcendental phenomenological perspective. From this perspective, the core of the Vedāntan argument concerns not so much episodic memory in the sense of the distinct mental act of recollection but rather what Husserl calls “retention”—the holding onto the just-past as an intentional content belonging to our consciousness of the passage of time, including our own mental lives as flowing in time. The Advaita Vedāntan thought is that, at the moment of waking up, I can experience by retentional awareness my having just been asleep and my having not known anything. What Nyāya fails to see, according to Vedānta, is that I need this kind of retentional awareness in order to have the first-person knowledge that I slept and to ground any retrospective inference I may subsequently make.

Of course, even if we suppose that there is or can be such a direct memory in the form of a retentional awareness of the deep sleep state, the presence of such a memory would not suffice to prove the continuous presence of consciousness throughout the entirety of dreamless sleep. After all, the presence of such a memory seems compatible with there having been moments or periods during which consciousness vanishes completely, with the sleeper remembering only the later smoothed-out and mentally-merged, conscious parts of sleep. Nevertheless, if dreamless sleep allows for or includes phases in which awareness is present, then this state cannot be defined as one in which consciousness is absent.

Another important Advaita Vedāntan thought is that when I say I just woke up from a dreamless sleep, the first-person pronoun does not refer to my autobiographical self—my self as I represent it in personal memory. Rather, it picks out my consciousness or subjectivity itself. To use a phenomenological idiom, it picks out the “ipseity” or minimal selfhood of consciousness in contrast to the ego as a mentally represented object of memory or reflection. But whereas the Advaitin takes this minimal selfhood to be a transcendental “witness consciousness” (Gupta 1998), it is open to us today to

maintain that it is my embodied self or bodily subjectivity, or what phenomenologists would call my “pre-personal lived body.” In this way, we may be able to remove the Advaita Vedāntan conception of dreamless sleep from its native metaphysical framework and graft it onto a naturalist conception of the embodied mind—a conception that should also appeal to the Cārvāka or naturalist school of Indian philosophy (see [Ganeri 2012](#), pp. 69–97), besides being tractable for cognitive science.

Cognitive science is also relevant to an interesting disagreement between Yoga and Advaita Vedānta concerning cognitive activity during dreamless sleep. Advaita Vedānta maintains that cognitive activity ceases during dreamless sleep and only consciousness remains, whereas Yoga maintains that cognitive activity continues during dreamless sleep (see [Dasgupta 1922](#), pp. 460–61). To understand this difference it is important to note that both traditions distinguish between consciousness, which is the self-luminous (reflexive) and passive witnessing awareness, and the mind, which is the intentional or object-grasping cognitive system. Moreover, in the Yoga view, the mind is material, and so is not different from the body (see [Schweizer 1993](#)). According to Yoga, deep sleep is a subtle or reduced state of the mind, specifically of the “inner sense” (*antaḥkaraṇa*), which includes both mental cognition (*manas*, which processes and integrates sensory material, and *buddhi*, which intellectually discriminates and judges) and the sense of ego (*ahaṃkāra*, the feeling, “I am”). Thus, for Yoga, cognitive activity, particularly the formation of memories, continues subliminally in deep sleep, and this process is physical or physiological. According to Advaita Vedānta, however, the mind, specifically the inner mental sense, shuts down entirely in deep sleep, leaving only the passive “witness consciousness” and the life processes of the body. If we set aside the question of consciousness and ask whether cognitive activity, specifically memory formation, occurs during deep sleep, the answer from cognitive science is unequivocal, for evidence from psychology and neuroscience indicates that memory processes are strongly present in deep sleep ([Diekelmann & Born 2010](#); [Walker 2009](#)).

These processes include both passive and active forms of memory consolidation (the strengthening of newly-acquired memories and the integration of them with older ones). Of course, this kind of memory consolidation is thought to occur in the absence of consciousness, so this evidence does not support the Yoga and Vedāntan view that consciousness continues in dreamless sleep. Nevertheless, the evidence does support the Yoga view that physiologically-instantiated cognitive processes continue in dreamless sleep, contrary to both Advaita Vedānta and Nyāya, which believe the mind shuts down in dreamless sleep.

The claim that mental activity ceases in dreamless sleep while consciousness remains creates another difficulty for the Advaita Vedāntan view. If the inner sense stops functioning in dreamless sleep, then how is the waking memory, “I slept peacefully and I did not know anything,” formed? Episodic memory requires the encoding of experience, so if there is no experience of “I” in dreamless sleep, then how can I remember that I slept well?

The Advaita Vedānta answer is clever (see [Dasgupta 1922](#), pp. 460–461). In deep and dreamless sleep, ignorance completely envelops the mind. Since the ego sense is inoperative, it doesn’t appropriate this ignorance to itself, so there is no feeling of the ignorance belonging to an “I.” At the moment of awakening, however, the ego sense, grounded on the felt presence of the body, reactivates, and the mind starts up its cognitive workings. Immediately, the ego sense appropriates the lingering impression or retention of not-knowing and associates this retention with itself, thereby generating the retrospective thought, “I did not know anything.”

From the Vedānta perspective, this “I” is not the true self; it consists in a mistaken superimposition of the self onto the mind-body complex. The true self is the egoless “witness consciousness” (egoless, because it is not a function of the ego sense). The Advaitin takes this “witness consciousness” to be transcendental and not essentially embodied. It is open to us today, however, to suppose that if there is some kind of egoless and basal consciousness that can continue to be present in dreamless sleep, then it is a funda-

mentally embodied consciousness, perhaps a minimal mode of sentience consisting in the feeling of being alive. This thought provides another example of how it may be possible to separate the Advaita Vedāntan conception of consciousness in dreamless sleep from its original metaphysical framework and graft it onto a contemporary naturalist conception of the embodied mind.

If we project some terminology from contemporary philosophy of mind onto Yoga and Advaita Vedānta, then we can say that dreamless sleep counts for these Indian philosophers as a “phenomenal state” or a state of “phenomenal consciousness”—a state that has a phenomenal character or for which there is something it is like to be in that state. What is it like? Yoga and Vedānta describe deep and dreamless sleep as peaceful, as one undifferentiated awareness not divided up into a sense of being a distinct subject aware of a distinct object, and as blissfully unknowing. From a contemporary naturalist perspective, this conception could be taken as a description of a quiescent and tranquil form of sentience or the feeling of being alive. Under this description, dreamless sleep would not count as a state of “access consciousness”—a state whose phenomenal content or character we can cognitively access, hold in working memory, and use to guide our attention and thinking. We seem to have no cognitive access to being asleep during sleep; rather, we gain access retrospectively in the waking state. On this conception, in dreamless sleep we are phenomenally aware but we have no cognitive access to that awareness at the time.

Ultimately, however, this way of conceptually parsing the Yoga and Vedāntan view will not work. A central commitment of Yoga and Vedānta, as well as Indo-Tibetan Buddhism, is that we can gain access to the state of dreamless sleep through meditative mental training. I will come back to this idea at the end of this paper. But first we need to consider the default view of consciousness and dreamless sleep in cognitive neuroscience.

4 Assessing the default view

Why have neuroscientists thought that consciousness disappears during dreamless sleep?

One reason comes from the reports that people give when they are woken up from NREM (non-Rapid Eye Movement) sleep, especially when the EEG shows slow waves in the delta frequency range (0.5–4 Hertz) during sleep stages 3 and 4 (so-called slow-wave sleep). When given the instruction, “report anything that was going through your mind just before waking up,” people tend to report short and fragmentary thoughts or not being able to remember anything at all (Nielsen 2000; Tononi & Koch 2008, p. 243). On the basis of such reports, scientists conclude that the sleepers were aware of little or nothing at all prior to being woken up, and hence that slow-wave sleep is a state of reduced or absent consciousness.

We need to be cautious here, however. The fact that one has no memory of some period of time does not necessarily imply that one lacked all consciousness during that time. One might have been conscious—in the sense of undergoing qualitative states or processes of sentience or awareness—but for one reason or another one was not able to form the kind of memories that later one can retrieve and verbally report.

This point is familiar to scientists who study the effects of anaesthetics (Alkire et al. 2008). At certain doses, some anaesthetics prevent memory formation while sparing awareness. Near the threshold of unconsciousness, some anaesthetics block working memory, but patients may still be aware and fail to respond because they immediately forget what to do. At lower doses, patients under general anaesthesia can sometimes carry on a conversation using hand signals, but after the operation they deny ever being awake.

Although dreamless sleep and anaesthesia are not the same condition, the general point that retrospective oblivion does not prove a prior lack of consciousness must be kept in mind whenever we are tempted to infer that consciousness is absent in deep sleep because people report not being able to remember anything when they are woken up.

We also need to think about the kinds of verbal reports that people are asked to make when they are woken up in the sleep lab. The instruction to report “anything going through

your mind just before waking up” encourages you to direct your attention and memory to the objects of your awareness—to anything you might have been thinking about. But what about the felt qualities or phenomenal character of your state of awareness? A different instruction would be to report “anything you were feeling just before waking up.” This instruction encourages you to direct your attention and memory to the felt quality of your sleep. Did you have any feeling of being aware? Was your sleep peaceful and clear, or was it agitated, restless, or sluggish? Or do you have no impression of any feeling or quality of awareness? The point here is to guide people away from focusing exclusively on the intentional objects of consciousness, which may be absent in deep sleep, and to orient them towards the felt qualities or phenomenal character of awareness itself.

Another reason neuroscientists think that consciousness fades away in deep sleep comes from comparing brain activity during slow-wave sleep with brain activity during waking consciousness. For example, during wakefulness, when an electrical pulse is used to stimulate a small region of the brain, the pulse generates an EEG response that lasts for 300 milliseconds and that is made up of rapidly changing waves that propagate in specific directions over long distances in the cortex (Massimini et al. 2005; Tononi & Massimini 2008). During deep sleep, however, although the initial EEG response to the stimulation is stronger than during wakefulness, the response remains localized to the stimulated region instead of travelling to distant regions, and it lasts only 150 milliseconds. In short, whereas the waking brain responds to stimulation with a complex pattern of large-scale activity across many interconnected regions, the deeply sleeping brain responds with localized and short-lived activity. These findings are interpreted as showing that “effective connectivity”—the ability of neural systems to influence each other—breaks down in deep sleep. As a result, “large-scale integration” (Varela et al. 2001) in the brain cannot happen—that is, the brain cannot generate the kinds of dynamically-changing large-scale patterns of activity that are known to characterize consciousness in the waking state.

But what is it about the loss of effective connectivity and large-scale integration that makes neuroscientists think that consciousness disappears in deep sleep? To put the question another way, what is the connection between the presence of consciousness and the presence of effective connectivity and large-scale integration?

To answer this question, neuroscientists usually rely on the idea that a content of consciousness is a reportable content, and that reportable contents are ones that can be attentionally selected, held in working memory, and used to guide thought and action. Such cognitive processes—selective attention, working memory, sequential thought, and action guidance—require the large-scale integration of brain activity.

One of the more theoretically-principled versions of this idea is Giulio Tononi’s “integrated information theory” of consciousness (2008). According to this theory, any typical conscious experience has two crucial properties. First, it is highly “informative,” in the technical sense that it rules out a huge number of alternative experiences. Even an apparently simple conscious experience, such as lying on your back and seeing the clear blue sky throughout your whole visual field, is richly informative in the sense that it rules out a vast number of other experiences you could have had at that moment. Second, the experience is highly “integrated,” in the sense that it cannot be subdivided into parts that you experience on their own, such as the top and bottom portions of your visual field, or the color and the space of the sky.

Given this model of consciousness as “integrated information,” Tononi proposes that the level of consciousness of a system at a given time is a matter of how many possible states (information) are available to the system as a whole (integration). In the waking state, many possible states are available to the whole system (the system is rich in integrated information), whereas in deep sleep this repertoire drastically shrinks to just a few states (the system is poor in integrated information). Transposed onto the brain, the idea is that during slow-wave sleep

there is a massive loss of integrated information in the brain. Effective connectivity breaks down, leaving isolated islands that cannot talk to each other (loss of integration), while the repertoire of possible states contracts to a few largely uniform states (loss of information). Hence, according to the integrated information model, deep sleep is a state where consciousness reduces to a very low level or disappears entirely.

Although the integrated information theory offers a useful way of thinking about the qualitative richness and coherence of consciousness in informational terms, the theory has serious limitations as a theory of phenomenal consciousness, so it would be a mistake to use the theory to rule out the possibility of consciousness during dreamless sleep.

Despite Tononi's bold claim that "consciousness is one and the same thing as integrated information" (2008, p. 232), integrated information does not seem sufficient for consciousness. On the one hand, even simple systems have some degree of integrated information, so the equation of consciousness and integrated information implies that even simple systems, such as a photodiode, have some degree of consciousness. On the other hand, complex digital computers can possess a high amount of integrated information. Yet neither system is conscious (at least the attribution of consciousness to such systems seems highly implausible) (see Searle 2013). As Ned Block (2009) points out, the integrated information theory fails to distinguish between intelligence, in the sense of being able to solve complex problems by integrating multiple sources of information, and consciousness, in the sense of sentience or felt awareness (phenomenal consciousness). Since integrated information does not seem sufficient for consciousness—let alone identical to it—the presence or absence of integrated information cannot be the crucial mark of whether a state is conscious or not conscious.

We also need to keep in mind the distinction between "phenomenal consciousness" and "access consciousness." To be phenomenally conscious means to be in a state that has some subjective or phenomenal character. To be ac-

cess conscious means to be in a state where there is cognitive access to the contents of awareness. Whether a state's being phenomenally conscious requires that it be cognitively accessible is currently a matter of debate (Block 2011; Cohen & Dennett 2011). Although large-scale integration in the cortex is crucial for cognitively accessed or reported conscious experience, it may not be crucial for every kind of phenomenal consciousness; for example, it may not be crucial for the kind of cognitively unaccessed consciousness that Yoga and Vedānta maintain is present in dreamless sleep (though they also maintain, as we shall see, that this kind of consciousness is accessible if one is highly trained in certain types of meditation).

The upshot of this critical assessment of the default view is that neither the subjective report data nor the objective neurophysiological data suffice to rule out the possibility of a subtle mode of phenomenal consciousness occurring in certain phases of dreamless sleep. To put the point another way, the sleep science construct of "dreamless sleep," defined electrophysiologically as slow-wave sleep, may need phenomenological refinement. We need to allow for the possibility that certain types of slow-wave sleep may have a phenomenal character—a possibility that could in turn lead to refinements in the physiological construct of slow-wave sleep. It follows from these considerations that the standard neuroscientific definition of consciousness as "that which disappears in dreamless sleep and reappears in waking and dreaming states" is not acceptable. At the very least, it needs qualification in light of the present considerations, and it may need to be either substantially revised or abandoned in light of further research.

The case of dreamless sleep suggests that we need to allow at least for the possibility of there being modes of phenomenal consciousness that may not be cognitively accessible in the usual ways. At the same time, Yoga and Vedānta, as well as Indo-Tibetan Buddhism, maintain that aspects of the mind in deep and dreamless sleep can become cognitively accessible through meditative mental training. This is the last topic I wish to discuss. My main point

will be that considering sleep from this contemplative angle suggests new experimental questions and protocols for the cognitive neuroscience of sleep and consciousness.

5 New experimental questions and protocols

In juxtaposing the Indian and neuroscience conceptions of deep sleep, I have proceeded so far as if the Indian notion of dreamless sleep corresponds to NREM slow-wave sleep. But we can now see that this correspondence is too simplistic. The Indian conception of dreamless sleep suggests that we need a finer taxonomy of sleep states—a taxonomy that is not just physiological but also phenomenological, and that accommodates the ways that sleep may be culturally variable as well as flexible and trainable through meditative practices.

Consider that the fourth century C.E. author, Vyāsa, in his commentary on Patañjali's *Yoga Sūtras*, distinguishes three types of sleep that are recalled upon awakening—peaceful sleep, disturbed sleep, and heavy sleep. According to the cosmology that informs Yoga, these three types of sleep result from whichever one of the three qualities or tendencies (*guṇas*) predominates in the psychophysical complex. Overall, the quality of dullness or the tendency to inactivity (*tamas*) dominates the mind in ordinary sleep. Sleep is heavy or stupefying when this quality is not modified by either of the two other qualities or tendencies. Sleep is disturbed and restless when the quality of excitation or tendency to activity (*rajas*) is present. And sleep is peaceful and refreshing when the quality of lightness or tendency to clarity (*sattva*) is present. When the Vedānta philosophers describe deep and dreamless sleep as blissful, it is deep sleep, with this quality of clarity, that they have in mind.

When sleep-lab participants are roused from NREM sleep, however, they sometimes report that they have been thinking while they were asleep, and often they describe going around in a repetitive loop of rumination. Although this kind of thinking probably occurs mainly in stage 2 NREM sleep, it is also repor-

ted during awakenings from deeper slow-wave sleep.

Owen Flanagan appeals to this finding to argue that there is no such thing as dreamless sleep and hence no sleep completely lacking in consciousness (2000). Contrary to the standard neuroscience view, Flanagan thinks we are always conscious while asleep because we are always dreaming. Dreaming, he proposes, is any conscious mental activity occurring during sleep, not just mental activity involving sensory imagery. If ruminative thinking occurring in NREM sleep counts as dreaming, and if this kind of mental activity can happen during slow-wave sleep, then all sleep stages involve dreaming and at least some degree of consciousness.

From the Indian perspective, however, we need to distinguish clearly between two things. One is whether there is such a thing as dreamless sleep; the other is whether we are conscious while we sleep. Yoga and Vedānta agree that consciousness is present while we sleep, but this is not because we are always dreaming, even if we define “dreaming” widely to mean any kind of thinking during sleep. On the contrary, what Yoga and Vedānta mean by “dreamless sleep,” as we have seen, is that sleep state in which there are no sensory or mental objects of awareness, that is, no images and no thoughts. Nevertheless, they maintain, there is awareness, so this state is a conscious state; it is a mode of consciousness without an object.

In the Yoga framework, reports of ruminative thinking upon awakening indicate a coarser or shallower sleep state—one closer to the surface of thinking consciousness—and a state with a strong quality of excitation or tendency toward movement of the mind.

Consider now the reasons that sleep scientist J. Allan Hobson gives for doubting the reliability of waking reports of ruminative thinking during slow-wave sleep:

Reports of antecedent mental activity elicited following awakenings from deep sleep are rendered unreliable by the brain fog through which they must pass [...]. Even if the deeply sleeping brain were capable of

the low-level ruminations sometimes implied by experimental reports, it is unlikely that they would survive the inertia of awakening. It may even be that the tumult of the awakening process triggers the chaotic and fragmentary mentation that is reported. And even when deep sleepers are sufficiently aroused to be interviewed, they may still generate huge slow waves in their EEGs, indicating that they are in a semituporous state quite different from either sleeping or waking. Indeed, they may even hallucinate, become anxious, and confabulate as if they suffered from delirium. This is precisely what happens in the night terrors of children. (1999, pp. 142–143)

Clearly, this too is a far cry from the Indian conception of dreamless sleep. Neither reports of ruminative thinking nor waking hallucinatory confabulations correspond to the Yoga and Vedāntan descriptions of dreamless sleep as a peaceful or blissful state free of mental activity, from which one awakens feeling alert and refreshed. From the Yoga perspective, what Hobson describes are sleep states strongly marked by a quality of dullness combined with mental excitation upon awakening.

My point here is not at all that sleep science should refine its taxonomy using the Yoga framework. It is rather that ultimately we cannot map the Indian notion of dreamless sleep using already-established scientific categories, especially the physiologically-defined sleep stages, which, even from a scientific perspective, are now recognized as too crude to capture the moment-to-moment dynamics of electrical brain activity during sleep, let alone the experiences with which they may be correlated (Nir & Tononi 2009). Not only is the Indian notion phenomenological and metaphysical, rather than physiological, it is also embedded in a normative framework that understands sleep in contemplative terms. So, to bridge from sleep science and the neuroscience of consciousness to the Indian conception of dreamless sleep, we need to view sleep as a mode of being that is trainable through meditation.

From the Yoga perspective, entering a state of blissful dreamless sleep on a regular basis requires leading a calm and peaceful life guided by the fundamental value of nonviolence (*ahimsā*), practicing daily meditation, and treating going to sleep and waking up as themselves occasions for meditation—for watching the mind as it enters and emerges from sleep.

In addition, from a yogic perspective, we need to distinguish between ordinary dreamless sleep and lucid dreamless sleep. Ordinary dreamless sleep is the sleep of ignorance, in which awareness is described as being in total darkness. Lucid dreamless sleep is described as a state in which awareness is luminous and without an object (free of thoughts and images). Whereas lucid dreaming consists in knowing that you are dreaming, lucid dreamless sleep is said to consist in being able to witness the state of dreamless sleep and recall its phenomenal clarity upon waking up. Although the background metaphysics of Yoga, Vedānta, and Indo-Tibetan Buddhism differ in significant ways, they all describe lucid dreamless sleep as disclosing a basal level of pre-personal consciousness that lies deeper than the modes of awareness that characterize the ego-centred waking and dreaming states.⁶

At this point you may wonder whether we have strayed back into the realm of metaphysics. Does this conception of dreamless sleep really have any descriptive phenomenological content or is it simply a consequence of the Indian metaphysical views that identify the true self with pure consciousness (as in the case of Vedānta) or that maintain that there is no self but only an ownerless stream of consciousness that continues in dreamless sleep (as in Indo-Tibetan Buddhism)?

From a purely textual perspective, the metaphysical and the phenomenological are thoroughly intertwined in the Indian discussions. From a cognitive science perspective, however, we can ask whether the idea of inducing lucid dreamless sleep through certain types of meditation is experimentally testable,

⁶ For further discussion, see Thompson (2014).

and, more generally, whether meditation is associated with altered sleep patterns or has measurable effects on sleep. Two neuroscience studies of sleep in relation to meditation are suggestive in this regard.

One recent study comes from the laboratories of Giulio Tononi and Richard Davidson (Ferrarelli et al. 2013). They examined slow-wave sleep in highly experienced Theravāda Buddhist and Tibetan Buddhist meditation practitioners. They found that the long-term meditators, compared to non-meditators, had significantly increased fast-frequency gamma activity, as recorded by high-density EEG, in a parietal-occipital region of the scalp during NREM sleep. In addition, the higher gamma activity was positively correlated with the length of meditation training. This finding is notable because gamma-frequency electrical brain activity is a well-known neural marker of conscious cognitive processes (Tononi & Koch 2008), including certain types of meditative states in long-term meditation practitioners (Lutz et al. 2004). Gamma activity has also been shown to distinguish lucid dreaming from non-lucid dreaming in REM sleep (Voss et al. 2009; see also Voss & Hobson this collection). During NREM sleep, however, gamma activity tends to decrease, so the higher gamma activity in the meditators could reflect a capacity to maintain some level of awareness. More generally, the study suggests that there may be distinct slow-wave sleep states associated with meditation practices.

Another older study examined long-term practitioners of TM (Transcendental Meditation) who reported what they called the subjective experience of “witnessing” during sleep (Mason et al. 1997). They described this experience as one of feeling a continuous and peaceful awareness without dreams while one sleeps and as resulting in one’s feeling refreshed upon awakening. The main finding was that the long-term meditation practitioners, compared to short-term practitioners and non-meditators, showed a unique EEG pattern during slow-wave sleep, one in which faster alpha and theta waves were superimposed on the slower delta waves. Although we cannot draw clear conclusions

about what these distinctive physiological patterns mean, including whether they are due to TM practice or some other cause, the authors of the study interpret them as supporting the presence of a different kind of slow-wave sleep state in individuals who report witnessing of sleep.

These two studies reinforce the point that we cannot use already established categories from sleep science to map the Indian conception of dreamless sleep. This conception, besides being closely tied to a specific phenomenology, which in turn reflects a specific metaphysics, is embedded in a normative cultural framework that aims to bring about and promote certain kinds of contemplative sleep states. Instead of trying to fit these states into a physiological scheme derived from studying the way twentieth-century Americans and Europeans sleep in the sleep lab, we need to enlarge the conceptual framework of sleep science to include contemplative ways of training the sleeping mind. This project will require that sleep scientists, cognitive neuroscientists, cognitive anthropologists, and Western and Indian philosophers work together to map the sleeping mind. In short, we need a cross-cultural cognitive science and neurophenomenology (Lutz & Thompson 2003) of the wake–sleep cycle, one that draws on the combined expertise of Western and Asian theoretical traditions.

One benefit of such a cross-cultural cognitive science is that it could offer new data relevant to our guiding question about consciousness and dreamless sleep. Consider the following testable, neurophenomenological hypothesis: In highly-experienced practitioners of certain types of meditation, compared to individuals without this kind of experience, we should observe a stronger correlation between subjective reports of phenomenal qualities of sleep and various objective measures of brain activity. Specifically, if highly experienced meditators were able to provide reports upon awakening about qualities of their experience of the state they call dreamless sleep, and if cognitive neuroscientists were able to relate these reports to fine-grained features of sleep physiology and to familiar aspects of the neural correlates of consciousness, then we would have new evidence

from experimental science that a certain type of dreamless sleep in certain individuals counts as a mode of phenomenal consciousness whose felt qualities can be made accessible to verbal report.⁷

This hypothesis also cast lights on our earlier discussion of sleep-state misperception. From a contemplative perspective, when little attention is given to sleep as an occasion for the practice of mindfulness, it is not surprising that sleep-state perception will be unreliable, even in ordinary individuals, let alone patients suffering from insomnia or other sleep disorders. In contrast, sleep-state perception may be more reliable when sleep is valued in a contemplative way and is treated as an opportunity for cultivating mindfulness. Whether these assumptions are correct is something that neurophenomenology should test.

6 Conclusion

The definition of consciousness as “that which disappears in dreamless sleep and reappears when we wake up or dream” is unsatisfactory. It rules out the possibility of states or phases of dreamless sleep in which some kind of consciousness is present. A strong case for taking seriously this possibility can be constructed by combining resources found in Indian philosophy, Western philosophy of mind, the neuroscience of consciousness, and sleep science. The main message of this paper—besides that of needing to revise the above definition of consciousness—is that we need a more refined taxonomy of sleep states than the one that sleep science and the neuroscience of consciousness currently employ, and that contemplative methods of mind training are relevant for advancing the neurophenomenology of sleep and consciousness.

References

- Alkire, M. T., Hudetz, A. G. & Tononi, G. (2008). Consciousness and anesthesia. *Science*, 322 (5903), 876-880. [10.1126/science.1149213](https://doi.org/10.1126/science.1149213)
- Arya, P. (1989). *Yoga-Sūtras of Patañjali with the exposition of Vyāsa. Volume I: Samādhi-pāda*. Honesdale, PA: The Himalayan International Institute.
- Baker, F. C., Maloney, S. & Driver, H. S. (1999). A comparison of subjective estimates of sleep with objective polysomnographic data in healthy men and women. *Journal of Psychosomatic Research*, 47 (4), 335-341. [10.1016/S0022-3999\(99\)00017-3](https://doi.org/10.1016/S0022-3999(99)00017-3)
- Block, N. (2009). Comparing the major theories of consciousness. In M. Gazzaniga (Ed.) *The cognitive neurosciences IV* (pp. 1111-1122). Cambridge, MA: MIT Press.
- (2011). Perceptual consciousness overflows cognitive access. *Trends in Cognitive Sciences*, 15 (12), 567-575. [10.1016/j.tics.2011.11.001](https://doi.org/10.1016/j.tics.2011.11.001)
- Bryant, E. F. (2009). *The Yoga Sutras of Patañjali*. New York, NY: North Point Press.
- Chapple, C. K. (2008). *Yoga and the Luminous: Patañjali's spiritual path to freedom*. Albany, NY: State University of New York Press.
- Cohen, M. A. & Dennett, D. C. (2011). Consciousness cannot be separated from function. *Trends in Cognitive Sciences*, 15 (8), 358-364. [10.1016/j.tics.2011.06.008](https://doi.org/10.1016/j.tics.2011.06.008)
- Dasgupta, S. (1922). *A history of indian philosophy*. Cambridge, UK: Cambridge University Press.
- Diekelmann, S. & Born, J. (2010). The memory function of sleep. *Nature Reviews Neuroscience*, 11 (2), 114-126. [10.1038/nrn2762](https://doi.org/10.1038/nrn2762)
- Ferrarelli, F., Smith, R., Denticò, D., Riedner, B. A., Zenning, C., Benca, R. M., Lutz, A., Davidson, R. J. & Tononi, G. (2013). Experienced mindfulness meditators exhibit higher parietal-occipital EEG gamma activity during NREM sleep. *PLoS ONE*, 8 (8), e73417-e73417. [10.1371/journal.pone.0073417](https://doi.org/10.1371/journal.pone.0073417)
- Flanagan, O. (2000). *Dreaming souls: Sleep, dreams, and the evolution of the conscious mind*. New York, NY: Oxford University Press.
- Ganeri, J. (2012). *The self: Naturalism, consciousness, and the first-person stance*. Oxford, UK: Oxford University Press.
- Gupta, B. (1995). *Perceiving in Advaita Vedānta: An epistemological analysis and interpretation*. Calcutta, IND: Motilal Banarsidass.
- (1998). *The disinterested witness: A fragment of Advaita Vedānta phenomenology*. Evanston, IL: Northwestern University Press.

⁷ For further discussion, see [Thompson \(2014\)](#).

- Hobson, J. A. (1999). *Consciousness*. New York, NY: Scientific American Library.
- Iyengar, B. K. S. (1996). *Light on the yoga sutras of Patañjali*. London, UK: Thorsons.
- Kesarcordi-Watson, I. (1981). An ancient Indian argument for what I am. *Journal of Indian Philosophy*, 9 (3), 259-272. [10.1007/BF00235382](#)
- Lutz, A., Greischar, L. L., Rawlings, N. B., Ricard, M. & Davidson, R. J. (2004). Long-term meditators self-induce high amplitude gamma synchrony during mental practice. *Proceedings of the National Academy of Sciences of the United States of America*, 101 (46), 16369-16373. [10.1073/pnas.0407401101](#)
- Lutz, A. & Thompson, E. (2003). Neurophenomenology: Integrating subjective experience and brain dynamics in the neuroscience of consciousness. *Journal of Consciousness Studies*, 10 (9-10), 31-52.
- Mason, L. I., Alexander, C. N., Travis, F. T., Marsh, G., Orme-Johnson, D. W., Gackenbach, J., Mason, D. C., Rainforth, M. & Walton, K. G. (1997). Electrophysiological correlates of higher states of consciousness during sleep in long-term practitioners of the Transcendental Meditation program. *Sleep*, 20 (2), 102-110.
- Massimini, M., Ferrarelli, F., Huber, R., Esser, S. K., Singh, H. & Tononi, G. (2005). Breakdown of cortical effective connectivity during sleep. *Science*, 309 (5744), 2228-2232. [10.1126/science.1117256](#)
- Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, 83 (4), 435-450. [10.2307/2183914](#)
- Nielsen, T. A. (2000). A review of mentation in REM and NREM sleep: 'Covert' REM sleep as a possible reconciliation of two opposing models. *Behavioral and Brain Sciences*, 23 (6), 851-866. [10.1017/s0140525x00974028](#)
- Nir, Y. & Tononi, G. (2009). Dreaming and the brain: From phenomenology to neurophysiology. *Trends in Cognitive Sciences*, 14 (2), 88-100. [10.1016/j.tics.2009.12.001](#)
- Perlis, M. L., Giles, D. E., Mendelson, W. B., Bootzin, R. R. & Wyatt, J. K. (1997). Psychophysiological insomnia: The behavioural model and a neurocognitive perspective. *Journal of Sleep Research*, 6 (3), 179-188. [10.1046/j.1365-2869.1997.00045.x](#)
- Phillips, S. (2009). *Yoga, karma, and rebirth: A brief history and philosophy*. New York, NY: Columbia University Press.
- Proust, M. (2003). *The way by Swann's*. London, UK: Penguin Books.
- Ram-Prasad, C. (2007). *Indian philosophy and the consequences of knowledge*. Burlington, VT: Ashgate Publishing.
- Raveh, D. (2008). Ayam aham asmīti: Self-consciousness and identity in the eighth chapter of the Chāndogya Upaniṣad vs. Śāṅkara's Bhāṣya. *Journal of Indian Philosophy*, 36 (2), 319-333. [10.1007/s10781-007-9031-7](#)
- Rosa, R. R. & Bonnet, M. H. (2000). Reported chronic insomnia is independent of poor sleep as measured by electroencephalography. *Psychosomatic Medicine*, 62 (4), 474-482.
- Schweizer, P. (1993). Mind/consciousness dualism in Sāṅkhya-Yoga philosophy. *Philosophy and Phenomenological Research*, 53 (4), 845-859. [10.2307/2108256](#)
- Searle, J. R. (2000). Consciousness. *Annual Review of Neuroscience*, 23 (1), 557-578. [10.1146/annurev.neuro.23.1.557](#)
- (2013). Can information theory explain consciousness? *New York Review of Books*, 60 (1)
- Sharma, R. K. (2001). Dreamless sleep and some related philosophical issues. *Philosophy East and West*, 51 (2), 210-231. [10.1353/pew.2001.0031](#)
- Thompson, E. (2014). *Waking, dreaming, being. Self and consciousness in neuroscience, meditation, and philosophy*. New York, NY: Columbia University Press.
- Tononi, G. (2008). Consciousness as integrated information: A provisional manifesto. *Biological Bulletin*, 215 (3), 216-242.
- Tononi, G. & Koch, C. (2008). The neural correlates of consciousness: An update. *Annals of the New York Academy of Sciences*, 1124, 239-261. [10.1196/annals.1440.004](#)
- Tononi, G. & Massimini, M. (2008). Why does consciousness fade in early sleep? *Annals of the New York Academy of Sciences*, 1129, 330-334. [10.1196/annals.1417.024](#)
- Varela, F. J., Lachaux, J.-P., Rodriguez, E. & Martinerie, J. (2001). The brainweb: Phase synchronization and large-scale integration. *Nature Reviews Neuroscience*, 2 (4), 229-239. [10.1038/35067550](#)
- Voss, U., Holzmann, R., Tuin, I. & Hobson, J. A. (2009). Lucid dreaming: A state of consciousness with features of both waking consciousness and non-lucid dreaming. *Sleep*, 32 (9), 1191-1200.
- Voss, U. & Hobson, A. (2014). What is the State-of-the-Art on Lucid Dreaming? In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1-20). Frankfurt a. M., GER: MIND Group.
- Walker, M. P. (2009). The role of sleep in cognition and emotion. *Annals of the New York Academy of Sciences*, 1156, 168-197. [10.1111/j.1749-6632.2009.04416.x](#)

- Windt, J. M. (2010). The immersive spatiotemporal hallucination model of dreaming. *Phenomenology and the Cognitive Sciences*, 9 (2), 295-316.
[10.1007/s11097-010-9163-1](https://doi.org/10.1007/s11097-010-9163-1)
- Windt, J. M. & Metzinger, T. (2007). The Philosophy of dreaming and self-consciousness: What happens to the experiential subject during the dream state? In P. McNamara & D. Barrett (Eds.) *The new science of dreaming, volume III: Cultural and theoretical perspectives on dreaming* (pp. 193-247). Westport, CT: Praeger.
- Zhang, L. & Zhao, Z. H. (2007). Objective and subjective measures for sleep disorders. *Neuroscience Bulletin*, 23 (4), 236-240. [10.1007/s12264-007-0035-9](https://doi.org/10.1007/s12264-007-0035-9)
- Āraṇya, Sāṃkhya-yogāchāra Swāmi Hariharānanda. (1983). *Yoga philosophy of Patañjali*. Albany, NY: State University of New York Press.

Just in Time—Dreamless Sleep Experience as Pure Subjective Temporality

A Commentary on Evan Thompson

Jennifer M. Windt

In this commentary, I propose a strategy for extending Evan Thompson's argument on the existence of dreamless sleep experience. My first aim is to show that the Indian debate on reports of having slept peacefully is importantly similar to debates in scientific dream research and contemporary Western philosophy on the trustworthiness of dream reports. This analogy leads to a surprising conclusion: the default view of conscious experience as that which disappears in dreamless sleep, though widely accepted in cognitive neuroscience, is in fact inconsistent with the methodological background assumptions of scientific dream research. Importantly, the methods already used in scientific dream research, as well as the theoretical justification on which they are based, can be extended to the investigation of dreamless sleep experience. Second, I sketch the outlines of a conceptual model of dreamless sleep experience as involving pure subjective temporality, or phenomenal experience characterized only by the phenomenal *now* and the sense of duration, but devoid of any further intentional content. I suggest that understood in this manner, dreamless sleep experience is a candidate for minimal phenomenal experience, or the simplest form in which a state can be phenomenally conscious. This model also extends existing work on minimal phenomenal selfhood in dreams. Third, I discuss three empirical examples that I take to be particularly promising candidates of dreamless sleep experience. These are certain forms of minimal or imageless lucid dreams, white dreams, and sleep-state misperception of the type most dramatically seen in subjective insomnia.

Keywords

Dreaming | Dreamless sleep | First-person reports | Insomnia | Lucidity | Minimal phenomenal experience | Minimal phenomenal selfhood | Sleep-state misperception | Time consciousness | White dreams

Commentator

Jennifer M. Windt
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

Target Author

Evan Thompson
evan.thompson@ubc.ca
University of British Columbia
Vancouver, BC, Canada

Editor

Thomas Metzinger
metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

1 Introduction

The default view in philosophy of mind and cognitive neuroscience has long been that the very notion of phenomenal experience occurring during dreamless sleep is nonsensical and involves a conceptual contradiction.¹ In this view, consciousness is „that which disappears in dreamless sleep and reappears when we wake up or dream” (Thompson 2015, this collection, p. 1), and dreamless sleep is simply characterized by the absence of conscious experience.² In his target article, Evan Thompson casts doubt on this view. Drawing from classical Indian philosophy as well as evidence from sleep and dream research, he argues that dreamless sleep experience is a theoretically coherent and empirically tractable target for future research. Yet, in order to even begin to make sense of dreamless sleep experience, a more fine-grained taxonomy

of sleep states and new experimental protocols integrating disciplined first-person reports as well as neuroscientific methods are needed.

Here, I take up this challenge and attempt to sketch the outlines of a positive account of dreamless sleep experience. This commentary has three main aims. The first is to propose that Thompson’s case for dreamless sleep experience can be strengthened by constructing a rough analogy between the historical Indian debate on dreamless sleep and contemporary Western debates from scientific dream research and philosophy on the epistemic status of dream reports. Based on this analogy, I argue that the default view is inconsistent with the methodological background assumptions of scientific sleep and dream research. This internal inconsistency lends additional urgency to Thompson’s demand for a more fine-grained taxonomy of sleep states. I then use the Indian debate as a foil to sketch the outlines of an integrated theoretical position on the trustworthiness of first-person reports of dreams and dreamless sleep experience. I take this approach to be in the spirit of the type of cross-cultural approach recommended by Thompson and hope to show that valuable lessons can be learned on both sides.

My second aim is to sketch the outlines of a positive account of dreamless sleep experience. Here, my key claim is that dreamless sleep experience can be described as pure temporal experience. By this I mean phenomenal states that aside from their temporal structure are devoid of any further intentional content and characterized only by the subjective experience of time. Pure temporal experience (or pure subjective temporality, as I will also sometimes call it) is not structured around perceptual objects, events or emotions; it is the experience of being *just in time*.³ This account of dreamless sleep

¹ In some readings of the term dreamless sleep, the default view is not just obviously false, but it is also unclear that it is actually endorsed by many researchers working on dreaming and sleep. Most would acknowledge, for instance, that hypnagogic imagery during sleep onset or repetitive and non-progressive types of sleep thinking involve phenomenal experience during sleep; yet, because they are also commonly distinguished from full-fledged dreaming, they can be said to occur in dreamless sleep. This, however, is different from the type of dreamless sleep experience that Evan Thompson has in mind and that is the focus of this commentary. As will become clear later, in the narrower reading endorsed by Thompson, dreamless sleep “is that sleep state in which there are no sensory or mental objects of awareness, that is, no images and no thoughts” (p. 14); the question, denied by the default view, is whether this state of sleep can sometimes involve phenomenal experience. Dreamless sleep experience of this type, if it exists, is also distinct from experiences occurring during sleep-wake transitions in that it is thought to occur during deep sleep. In the context of this commentary, I will always, unless explicitly noted otherwise, use the term dreamless sleep experience in this narrow sense. In other readings, the default view may be thought to be trivially true: if one defines dreams as involving *any* kind of phenomenal experience during sleep (Flanagan 2001), then the occurrence of phenomenal experience during dreamless sleep is indeed ruled out by conceptual considerations. This reading, however, is too permissive in that it fails to acknowledge the distinction between different types of experiences occurring during sleep, ranging from imagistic, narratively complex, and often emotional dreams to thought-like activity. For now, this suggests that the default view is too simple: the question is not whether there are experiences during sleep that fall short of full-fledged dreaming in some particular sense but whether there is a further group of experiences—call them dreamless sleep experience in the narrow sense—that is distinct from *any* of the established forms of conscious experience during sleep, including hypnagogic imagery and sleep-thinking. Thompson acknowledges this issue (p. 14) and I only emphasize it here to avoid misunderstanding.

² Note that throughout this commentary, I will use the terms “experience”, “subjective experience”, and “consciousness” interchangeably to describe states that have phenomenal character, or for which there is something it is like to have them.

³ At first sight, there is an inherent ambiguity in the concept of pure subjective temporality in that it can refer to the experiential character of *nowness*, but also to the experience of duration and of succession. In section 4, it will become clear that in the account defended here, the two aspects of *nowness* and duration are not strictly dissociable: the simplest forms of temporal experience are characterized by both a phenomenal *now* and the experience of duration, because

experience is attractive, or so I claim, because it offers a way of spelling out not just what is distinctive about dreamless sleep experience, but also how dreamless sleep experience can be integrated into a broader theoretical framework describing different kinds of sleep experiences, including dreams. The key idea is that while even the simplest forms of dreaming are characterized by phenomenal selfhood, or the experience of being or having a self, the transition from dreaming to dreamless sleep experience occurs when even this minimal form of phenomenal selfhood is lost. While the analysis of dreaming can help identify the conditions for minimal phenomenal selfhood, the analysis of dreamless sleep experience may provide a glimpse of an even simpler (and perhaps even minimal) form of phenomenal experience. In the final part of the commentary, I identify what I take to be the three most promising candidates for a future research program on dreamless sleep experience. These are lucid dreamless sleep, white dreams, and sleep-state misperception of the type most commonly seen in subjective insomnia. These examples broaden the scope of the target phenomenon by suggesting that the theoretical and experimental investigation of dreamless sleep experience extends beyond the case of expert meditators discussed by Thompson.

2 From the classical Indian debate to a new taxonomy of experience during dreamless sleep

In *Dreamless Sleep, the Embodied Mind, and Consciousness*, Evan Thompson retraces the steps of the classical Indian debate between the Advaitins and the Nyāyīyikas on the occurrence of conscious experience during dreamless sleep (see also Thompson 2014, chap. 8). The classical Indian debate is important, according to Thompson, because if the Advaita Vedānta

and Yoga claims about the persistence of consciousness during dreamless sleep are correct, the default view of consciousness as that which disappears during dreamless sleep is false and requires revision. In this section, I briefly reconstruct Thompson's main arguments and sharpen the precise points of agreement and disagreement in the classical Indian debate, as well as their overlap with questions raised in cognitive science and contemporary philosophy of mind. I also introduce three challenges to Thompson's view.

Thompson's reconstruction of the classical Indian debate starts out from a deceptively simple question: How, after awakening from sleep, do we know that we have slept peacefully? The Yoga and Advaita Vedānta schools argue that retrospective reports of having slept peacefully are memory reports: we directly and non-inferentially remember (and hence are able to report) a state in which we were phenomenally conscious, but did not experience any particular thoughts or images. Dreamless sleep experience is, in this view, devoid of intentional content; it is a state of knowing nothing and at least in principle, it can be remembered and accurately reported upon awakening. The Nyāyas disagree, arguing that reports of having slept peacefully are inferential. Their point is that if dreamless sleep involves a particular form of ignorance, or of not-knowing, this not-knowing cannot itself be known, either during sleep or retrospectively. Because the means for knowledge are lacking during dreamless sleep, we can at best infer, when we wake up feeling refreshed and remember nothing, that we must have slept peacefully.

As Thompson (sec. 3) points out, the classical debate about conscious experience during dreamless sleep has to be seen in the larger context of how these schools construe the relationship between consciousness and the self. For the Nyāyas, consciousness is an adventitious property of the self, meaning that the self can persist throughout sleep even when consciousness ceases. They also claim that cognition always involves taking something as its object, where this object is necessarily distinct from the cognitive state itself. This view is compatible with

the phenomenal *now* itself is temporally stretched. Though for reasons of space, I cannot discuss this any further here, note that once the distinction between the phenomenal *now* and the experience of duration collapses, the experience of seriality or of succession disappears as well: if the phenomenal *now* is no longer embedded within a larger temporal reference frame, then there will be no separate events that can be experienced as succeeding each other.

the occurrence of object-directed thought and dream-related imagery during sleep, but prohibits the occurrence of objectless cognitive states. For the Advaitins, the situation is different. Because for them, the self is pure, reflexive (or self-luminous) consciousness, they cannot allow that consciousness can disappear altogether even during sleep, because this would entail a disappearance of the self. Unlike the Nyāyas, the Advaitins do not, however, take consciousness to be necessarily object-directed. Instead, they regard the essentially reflexive and self-luminous character of consciousness and the self as separate from and indeed as the very condition of object-directed thought. A prediction would be that “pure” cases of reflexive, self-luminous consciousness should occur even in the absence of object-directedness, for instance during sleep.

Despite these differences, the debate on dreamless sleep experience unfolds before a background of mutual agreement. Both schools agree, for instance, that object-directed consciousness can (and does, for instance in the form of dreams) occur during sleep, but also that it does not persist throughout sleep. Both also agree that dreamless sleep is a state in which object-directedness is lost. And finally, both agree that the self persists throughout dreamless sleep, even in the absence of object-directedness. Their disagreement thus hinges, first, on what exactly it means to say that the self persists during dreamless sleep, understood in the sense of a state in which object-directed thought is lost, and second, on how to construe the relationship between consciousness, the self, and memory reports. Both points are relevant, as we will see, for assessing the relationship between the Indian debate and contemporary research as well.

How, then, to adjudicate between the two sides in the debate? Thompson (p. 6) reconstructs the Nyāya claim that our knowledge of dreamless sleep is inferential as involving a five-step syllogism. His discussion of the Nyāya syllogism is already so clear that nothing would be gained from rehearsing it once more here. Instead, I want only to recall to readers’ attention that Thompson’s reconstruction of the Advaitin re-

sponse shows the Nyāya syllogism to be inherently fallacious: it is either circular or results in an infinite regress. In order to infer from the fact that I was in a special state that I knew nothing in this state, I must first have a reason for saying that I was indeed in a special state; and if this reason is that I knew nothing in this state, I am presuming what is supposed to be shown and the argument is circular. Alternatively, if I say that the means for knowledge were lacking in this special state, for instance because the mental faculties and the senses were inactive, then this further claim has to be backed up by independent evidence. Saying that I felt refreshed upon awakening will not do—for in order to know that feeling refreshed after awakening is correlated with the inactivity of the mental faculties and the senses during sleep, I would either once more have to appeal to memory (which, on pains of circularity, I cannot do), or I would be headed for an infinite regress. Thompson sums up his critique of the Nyāya syllogism by formulating a general principle:

More generally, the only way I can know that the means for knowledge were absent in deep sleep is by knowing that there was no knowledge present in this state. Only by knowing the effect—my not knowing anything—can I infer the cause—the absence of the means for knowledge. So unless I already know what the inference is trying to establish—that I knew nothing—I cannot establish the reason on which the inference relies. (p. 7)

The Advaitin view offers an easy way out. As Thompson points out, it can be reconstructed as involving the phenomenological claim

that when I wake up from a dreamless sleep, it seems that I can sometimes knowingly say I have just emerged from a dreamless sleep, and this saying seems to be a reporting of my awareness, not the product of having to reason things out. (p. 8)

At least in principle, the subjective impression of having awakened from dreamless sleep can be

reflected in veridical reports of awareness during dreamless sleep.

It is important to see that Thompson's assessment of the Indian debate does not lead to a whole-hearted endorsement of the Advaitin view; the view he promotes is in fact much more subtle, and also more humble. Thompson's main goal is to establish the logical possibility of dreamless sleep experience. For this, it is sufficient that veridical memories of having slept dreamlessly are possible in principle (p. 5, p. 9). He also explicitly allows that there could be cases in which one's memory of having slept peacefully and dreamlessly is mistaken. Thompson's view is also weaker than the Advaitin position in that it is not committed to the persistence of conscious experience throughout sleep, but leaves room for periods of unconsciousness during sleep. According to Thompson, the mere possibility of dreamless sleep experience challenges the default view and highlights the need for a refined taxonomy of sleep states, because such a refined taxonomy is the condition for investigating dreamless sleep experience experimentally (p. 3).

To be sure, Thompson also offers some factual evidence for thinking that dreamless sleep experience actually exists: experienced meditators report witnessing or becoming lucid during dreamless sleep, and they show a changed pattern of EEG activity during slow wave sleep. Meditative training may, as Thompson suggests, facilitate cognitive access to the state of dreamless sleep (p. 11) and with it, more accurate reports. But his main point is that conceptual and empirical questions about dreamless sleep experience are well worth asking and that in order to do so, prominent theories of sleep, but also of consciousness (such as Tononi's *Integrated Information Theory*; see [Tononi 2008](#)) should at least make room for the possibility of its occurrence and require revision.

While I find Thompson's case for the logical possibility and conceptual coherence of dreamless sleep experience compelling, I worry that its humility makes it vulnerable to three related objections. A proponent of the default view could acknowledge that veridical reports of dreamless sleep experience are logically possible

but could insist that unless such veridical reports are identifiable and can be distinguished from nonveridical ones, such reports cannot be used for the experimental investigation of dreamless sleep experience, or only in a very small and admittedly special group of highly trained subjects. Thompson's own suggestions for the future investigation of dreamless sleep experience assume that this basic problem has been solved. For instance, he proposes that because dreamless sleep experience is supposed to be devoid of intentional objects, asking participants to report anything that was going through their minds before awakening, which is a question about the objects of awareness or the contents of consciousness, might be poorly suited to the target phenomenon. A good alternative, he suggests, would be to direct participants' attention to the phenomenal character of sleep itself, for instance by asking them to report any feelings or any qualitative states experienced before awakening (p. 12). Here, the proponent of the default view might object that this strategy falls short of a methodology for investigating dreamless sleep experience: In order to use reports of dreamless sleep experience as evidence, some rationale for distinguishing veridical reports from nonveridical ones is needed. Without this, the large-scale revision of standard sleep-state taxonomy demanded by Thompson may seem premature; Thompson's case for the mere possibility of dreamless sleep experience lacks the empirical grounding and research methodology to justify such a move.

A related problem is that in order to empirically investigate the occurrence of dreamless sleep experience, it is not enough to identify veridical reports of such experiences and distinguish them from nonveridical ones. Instead, in order to determine the frequency of dreamless sleep experience, one has to determine whether subjects can reliably report not just the presence of dreamless sleep experience, but also its absence. This problem is especially pronounced because Thompson's claim is not that experience persists throughout sleep. As we saw earlier, his view departs from the Advaitin claim in that he thinks that dreamless sleep experience occurs only occasionally and contrasts

with periods of genuine unconsciousness during sleep. A report-based methodology for investigating dreamless sleep experience will consequently have to assume not only that reports of dreamless sleep experience reliably indicate the presence of such experience during the preceding sleep period, but also that the absence of such experiences can be reliably reported, or at least that it can be inferred from the absence of reports of dreamless sleep experience. Unless this second condition is fulfilled, reports of dreamless sleep experience could be highly reliable in that they occur only when dreamless sleep experience was in fact present during the preceding sleep period, but could nonetheless fail to be sensitive to its actual frequency, for instance by only following a small proportion of such sleep experiences (for a discussion of the reliability and sensitivity of first-person reports, see [Fink unpublished manuscript](#)).

Thompson himself shies away from both commitments. In fact, he casts doubt on the assumption, common in cognitive neuroscience, “that a content of consciousness is a reportable content, and that reportable contents are ones that can be attentionally selected, held in working memory, and used to guide thought and action” (p. 12). Relatedly,

the general point that retrospective oblivion does not prove a prior lack of consciousness must be kept in mind whenever we are tempted to infer that consciousness is absent in deep sleep because people report not being able to remember anything when they are woken up. (p. 11)

Here, he might be read as effectively denying the possibility of using retrospective reports as a source of evidence for the scientific investigation of dreamless sleep experience. Moreover, given these doubts about the reliability and sensitivity of retrospective reports, Thompson’s (p. 17) proposal that meditation makes positive occurrences of dreamless sleep experience accessible to verbal report is not enough; a proponent of the default view could object that expertise of the relevant type is acquired only if meditation enables periods of unconscious sleep

to be retrospectively reported as well (or at least to be measured indirectly through the inability to report conscious experiences from the preceding sleep period).

Finally, a proponent of the default view might grant that reports of expert meditators are more trustworthy than those of laypeople in both respects: meditators can report both when dreamless sleep experience was present and when it was absent.⁴ Yet, it could still be objected that the example of expert meditators is simply too remote to justify the large-scale revision of sleep-state taxonomy that Thompson has in mind. For all practical purposes, or so the objection might go, the default view of consciousness and dreamless sleep as diametrically opposing and mutually exclusive states stands.

To be clear, I do not think these objections are particularly worrisome; but I do think they help set the agenda for how best to develop Thompson’s view, defend it against skeptical objections, and place it on broader empirical grounding. The first step, taken in the next section, is to introduce a stronger defense of the trustworthiness of reports of dreamless sleep experience, as well of reports of its absence. If successful, this provides a sound methodological basis for the experimental investigation of dreamless sleep experience. The second step is to provide a broader theoretical and empirical basis by proposing a conceptual framework of dreamless sleep experience as well as additional candidates for its future investigation.

3 Are reports of dreamless sleep experience trustworthy? The analogy between the Indian debate on dreamless sleep and the contemporary debate on dream reports

In this section, I draw an analogy between the Indian debate on dreamless sleep experience and the contemporary debate on the trustwor-

⁴ It remains controversial whether different forms of meditation actually enhance introspective accuracy. While there is some evidence in support of this claim ([Fox et al. 2012](#); [Sze et al. 2010](#)), at least one study has suggested that meditators may feel more confident than controls about their ability to successfully perform interoceptive tasks (such as heartbeat detection), but that this confidence is not paralleled by an actual improvement in task performance ([Khalsa et al. 2008](#)).

thiness of dream reports. This analogy provides the resources for overcoming the first two challenges to Thompson's argument. In particular, it reveals the default view to be inconsistent with the methodological background assumptions of scientific sleep and dream research. Given their own methodological commitments, researchers in these fields should reject the default view.

3.1 The methodological background assumptions of scientific dream research: Lessons for the investigation of dreamless sleep experience

The first step towards seeing why the default view is inconsistent with scientific dream research is to realize that this field, at least implicitly, relies on the assumption that reports of conscious experience during sleep are trustworthy: at least when they are given under certain (sufficiently) ideal conditions and immediately after awakening from sleep, such reports are taken to reflect what was experienced during the preceding sleep period, and indeed whether anything was experienced at all. What exactly the (sufficiently) ideal conditions for reporting sleep experiences consist in is an empirical question, and in scientific dream research, much work has been dedicated to its investigation (for discussion and further references, see Windt 2013, 2015, chaps. 3 and 4). There is widespread agreement that temporal proximity is a crucial factor: reports given immediately after awakening are commonly taken to be least vulnerable to forgetting. The sleeping environment (at home versus in the laboratory), method of awakening, interaction with experimenters, and precise wording of questions also play an important role (Domhoff 1996, 2003; Hall & Van de Castle 1966; Kramer 2013; Winget 1979). Different reporting techniques may be suitable for different research questions, and aside from being asked for verbal reports, participants may be encouraged to produce a dream drawing or compare the visual imagery in their dream with photographs with varying degrees, for instance, of color saturation or brightness (Rechtschaffen & Buchignani 1992). While there

may be uncertainty, in a given case, as to the sincerity of a report, this is a practical matter, not a deep theoretical problem.⁵ The key idea is that by improving reporting conditions and tailoring the reporting technique used in a given study to the specific research question, this risk can be minimized. For now, my main point is that this strategy, which is already well established in scientific sleep and dream research, only makes sense against a background of basic trust in at least a subset of dream reports.

This basic idea is very much in keeping with Thompson's proposal of asking participants to report any feelings or qualitative states experienced prior to awakening, rather than asking them to focus on the contents of conscious thought. By directing participants' attention to certain aspects of sleep experience or even introducing new experiential categories for their description (an excellent example of this strategy is Lutz et al. 2002), the expressive granularity⁶ of individual reports can be increased: types of experiences can be rendered reportable that would otherwise be forgotten. A compelling possibility is that in the case of dreamless sleep experience, such improvements in reporting conditions may not just supplement training, as suggested by Thompson, but may even facilitate the investigation of dreamless sleep experience in participants who lack any particular introspective training.⁷

Admittedly, this approach does not provide a fail-safe method for avoiding or even identifying nonveridical reports. Rather than fo-

⁵ Researchers occasionally worry, for instance, that participants may underreport embarrassing dream content; censorship of this type may be why sexual dream content is only rarely reported in laboratory studies (Hobson 1988); see also Rosen's (2013) discussion of willful narrative fabrication of dream reports. For the investigation of dreamless sleep experience, which is, after all, thought to be devoid of such content, such worries about censorship do not seem to apply.

⁶ I owe this term to Sascha Fink; see for instance Fink 2015, p. 23; for discussion, see Windt 2015, p. 92.

⁷ As Solomonova et al. (2014) note, it is important to distinguish questions about the range of possible experiences in dreams (or the "depth" of dreaming) from those about their typical characteristics in the general population (or the "breadth" of dreaming), and what counts as the ideal reporting conditions in the context of a given study depends on which of these questions is being addressed. For now, note that because expertise is likely most useful for answering questions about the depth of experience, and because expert reports may not be representative of the breadth of the target phenomenon, broadening the investigation of dreamless sleep experience beyond expert groups is an important goal for future research.

cusing on the veridicality of individual reports, the strategy is to identify which types of reports are best tailored to a given question and under which conditions they are most likely to be obtained. The problem of identifying individual reports of a certain type for which this strategy has failed is thus not obliterated, but minimized.⁸ What is more important is that there is, in this view, a distinction to be drawn between *general opinions* about experience and *reports of individual experiences*. Note that reports, in this context, are broadly construed as the product of (verbal or nonverbal) behaviors con-

ducted with the *sincere intent of conveying or recording certain relevant information about a specific dream* (for details, see Windt 2015, chap. 3.3) Questionnaires asking participants to assess the general frequency with which, for instance, they dream in color do not count as experience reports in this narrow sense. Indeed, there are good reasons for doubting the trustworthiness of responses to such general questionnaires, and in some cases, they have even been shown to be at odds with individual reports (Schwitzgebel 2002, 2011, chap. 1; Windt 2013, 2015, chap. 4.3). At best, such general questionnaires tap into opinions about experience, but whether these opinions match the phenomenal character of the corresponding experiences is a separate question. Importantly, questions about the relative trustworthiness of responses to general questionnaires can be meaningfully investigated only if the trustworthiness of at least a subgroup of dream reports is assumed (Windt 2015, chap. 4.4). This subgroup can then act as a baseline and can be used to determine the relative trustworthiness of answers to general questionnaires, but also of different types of reports. While the exact details continue to be debated (for instance on the laboratory effect), there is widespread agreement in scientific dream research that dream reports gathered immediately upon awakening, as is common in laboratory studies using timed awakenings from different sleep stages, are the gold standard against which other types of dream reports (such as home dream diaries compiled following spontaneous awakening) can be measured (again the debate on dream color is a good example; see Hoss 2010; Murzyn 2008; Schredl et al. 2008).

Importantly, as discussed earlier, the assumption that dream reports are trustworthy translates into a research strategy only if reports of nondreaming are taken to be equally trustworthy as reports of dreaming, at least when they are gathered under the same conditions. If the reporting conditions used in a given study are (sufficiently) ideal, it would, surely, be arbitrary to disqualify a subset of these reports on the basis of their content alone. In order to do so, some independent reasons for at-

⁸ Strictly speaking, it cannot be ruled out that even for reports obtained under seemingly ideal conditions—for instance immediately after awakening, and using appropriately worded questions—certain subject groups are particularly prone to memory failure or confabulation (Rosen 2013), or that results are distorted because of further disturbing factors that have so far been overlooked. The challenge will then be to identify such potentially disturbing factors, manipulate them experimentally, and derive certain predictions on how they will affect data obtained from the analysis of dream reports. These factors can then be integrated into a future, improved and more empirically plausible account of the ideal conditions of dream reporting. For now, my main point is that this strategy only makes sense if one already assumes that some subset of dream reports can be used as a baseline against which other, less trustworthy ones can be measured. The study of dream emotions is a nice example of how this strategy has been put to work in dream research. Views on the both the frequency and the types of emotions experienced in dreams have changed quite dramatically as new methods of collecting and scoring dream reports have been developed. Whereas older studies using classical dream content analysis suggested that dream emotions are relatively rare (Hall & Van de Castle 1966), the frequency of reported dream emotions increases tenfold when subjects are specifically asked to report their emotions on a line-by-line basis (Merritt et al. 1994). Affirmative probes of this sort suggest that dreams are “hyperemotional”, with emotions being mentioned in 95 percent of dream reports and the average dream report containing several different types of emotions. A plausible explanation is that dream emotions are underreported in free dream reports of the type used in older studies; free dream reports, in this view, are insufficient to capture the actual frequency of dream emotions. Until very recently, the accepted view was that the types of emotions experienced in dreams differ from those experienced in wakefulness in that dream emotions are predominately negative (Hobson et al. 2000). However, a recent study compared external ratings of emotions in dream reports to scores obtained when participants answered a standard emotion questionnaire themselves. Sikka et al. (2014) found that external ratings underestimate not only the frequency but also the types of emotions experienced in dreams. A particularly surprising result was that self-ratings showed positive dream emotions to be six times more frequent than negative ones. The systematicity of the differences is compelling and the same pattern was found in a number of follow-up studies (Sikka et al. 2014), suggesting that the use of self-ratings is a more reliable method for capturing the frequency and types of dream emotions than the use of external raters. This is not so say that the conditions for reporting and scoring dream emotions cannot be further improved. But this example does illustrate that theoretical views on dream emotions changed in tandem with changed and likely improved reporting and scoring conditions. Again, the idea is that methodological adjustments can obscure or render visible different aspects of the phenomenology of dreaming.

tributing reports of nondreaming to disturbing factors would be needed. It does not make sense to trust dream reports, but selectively distrust reports of nondreaming gathered under the same conditions and in the absence of any empirical evidence for distrusting them. Put differently, dreams will have to be regarded as reportable experiences, in the sense that given sufficiently ideal reporting conditions, their presence or absence, respectively, can actually be reported. Importantly, both assumptions are implicit in the scientific investigation of dreams. A brief excursion into the history of philosophical and scientific theorizing about sleep and dreaming illustrates this point.

The beginning of scientific dream research coincided with a new experimental paradigm: the practice of obtaining polysomnographic measurements of EEG activity, muscle tone and eye movements from subjects sleeping in the sleep laboratory and of obtaining mentation reports following timed awakenings. This methodology revealed reports of dreaming to be most frequent following awakenings from REM (rapid eye movement) sleep, whereas awakenings from NREM (non-REM) sleep were typically followed by an inability to recall any dreams. In their groundbreaking paper on the correlation between dreaming and REM sleep, [Aserinsky & Kleitman \(1953\)](#) optimistically claimed that that the method of timed awakenings from REM sleep “furnishes the means of determining the *incidence and duration of periods of dreaming*” ([Aserinsky & Kleitman 1953](#), p. 274; my emphasis).⁹ They very naturally took the reports given by their subjects to reflect conscious experience during the preceding sleep period, noting that “of 27 interrogations during [sic] ocular motility, 20 revealed detailed dreams usually involving visual imagery” ([Aserinsky & Kleitman 1953](#), p. 273; my emphasis). Because the method of obtaining reports following timed awakenings in the laboratory is, arguably, the backbone of scientific dream research, this assumption is not unique to Aserinsky and Kleit-

man’s original study. Instead, scientific dream research generally relies on the assumption that dream reports (at least when gathered under ideal reporting conditions, of which timed awakenings in the laboratory are taken to be a prime example) are epistemically transparent in the sense that they are trustworthy sources of evidence about the occurrence and phenomenal character of experience during sleep. I call this the *transparency assumption* ([Windt 2013, 2015](#)).¹⁰

It is important to see that on its own, the transparency assumption would be insufficient to establish the presumed correlation between dreaming and REM sleep. Claims about the sleep-stage or neural correlates of dreaming require that reports of dreaming and of nondreaming, when gathered under the same conditions, are equally trustworthy: if only reports of dreaming were trustworthy, but reports of nondreaming were not, then the analysis of dream reports would be insufficient to determine the occurrence and frequency of dreams during different sleep stages. Saying that dream reports are transparent is not quite enough: one will also have to assume that dreams are reportable experiences in the sense that had any dream occurred in a given sleep stage, one would in fact be able to report it, at least under sufficiently ideal reporting conditions. I call this the *reportability assumption* (for details, see [Windt 2015](#), chap.s 3 and 4). Only this added assumption casts reports of dreaming and of nondreaming as equally trustworthy and thus enables reports to be indicative of the occurrence and frequency of dreaming in different sleep stages. The emerging picture is that scientific dream research not just uses dream reports, *under the assumption of transparency*, to investigate *conscious experience during sleep*, but that in doing so, it is also *methodologically constrained by the space of reportable dreams*. Its implicit commitment to the

¹⁰ Here, I use the concept of epistemic transparency in a non-technical and metaphorical sense, intending to capture the intuition that dream reports are the closest researchers can come to “watching the sleeping mind” ([Cartwright 2010](#), p. 17). The choice of terminology also reflects the fact that dream reports are not identical with, but better conceived of as separate from dreaming. Finally, transparency is a nod to the historical situation that the theoretical problems raised by dream reporting were nearly invisible throughout most of the history of philosophical theorizing about dreaming.

⁹ Today, it is widely recognized that dreams can occur in all stages of sleep and are not exclusively a REM sleep phenomenon. Incidentally, this recognition may also lead to refined sleep-stage scoring systems and a blurring of the borders between REM and NREM sleep ([Nielsen 2000](#); see also [Windt 2015](#), chap. 2).

trustworthiness of reports of dreaming and of nondreaming means that it cannot go beyond what is in fact reported without risking internal inconsistency; it can only strive to render further aspects of dreaming reportable. Metaphorically speaking, the space of reportable dreams can be expanded; it can be broadened to cover more aspects of what characterizes typical dreams, or perhaps also to include more diverse types of dreams; and it can be deepened, by probing the unique aspects of certain types of dreams (such as nightmares) or the dreams of certain subject groups (such as meditators) in more detail (see [Solomonova et al. 2014](#)). Importantly, this reliance on dream reports is not a liability, a problem to be overcome: it is built into the very nature of dream research. Conversely, studies relying only on the polysomnographic analysis of sleep stages and/or neuroimaging data gathered independently of dream reports do not form part of dream research proper ([Windt 2015](#), chap. 3.2).

How does this account of dream reporting help address the objections to Thompson's argument discussed at the end of the last section? The strategy of focusing on reports gathered under (sufficiently) ideal reporting conditions and working towards a continuous improvement of these conditions is clearly relevant to the first objection, according to which the mere possibility of veridical reports is not enough. As soon as we broaden our focus from reports of dreamless sleep experience to reports of sleep experience (including dreams) more generally, it becomes clear that scientific dream research has long been centered on the project of identifying and optimizing the trustworthiness of such reports, as well as on determining the adequacy of different kinds of reports for addressing various research questions. Indeed, the very existence of scientific dream research hinges on the assumption that this can be done. Moreover, we have seen that the assumption that reports of dreaming and of nondreaming are equally trustworthy is implicit in this research strategy. This assumption is directly relevant to the second objection, according to which reports of dreamless sleep experience can be

used for the investigation of dreamless sleep experience only if they help detect both its presence and its absence.

Moreover, this proposal is, I think, compatible with Thompson's own strategy of focusing on reports from certain expert groups and improving the wording of questions. Indeed, this strategy of directing participants' attention to certain aspects of their experience rather than asking for a free report nicely parallels recent work suggesting that a self-scoring method, where participants respond to a standard questionnaire, for instance, about the emotions experienced in a particular dream, is a better measure of dream emotions than data obtained by external raters scoring free dream reports ([Sikka et al. 2014](#); see footnote 8 for discussion). This suggests that Thompson does not mean to reject, as a matter of principle, the claims that conscious experiences are reportable and that an absence of memory is sufficient to infer an absence of experience. Rather, I think his position involves the weaker claim that we should not easily and uncritically trust just any type of experience report to actually reflect the presence of such experience, nor should we easily and uncritically trust just any failure to remember previous experience as indicating an absence of such experience. But this weaker position is in keeping with the account of dream reporting outlined in this section. The challenge then becomes how to narrow the gap between experiences that are in fact reported and those that could (and would) be reported, given sufficiently ideal conditions. I think this is exactly the problem that large parts of report-based dream research are already trying to address.

Note that nothing I have said so far suggests that the transparency and reportability assumptions are theoretically justified (but see [Windt 2013, 2015](#)); if my analysis is correct, however, both are implicit in and in fact crucial for the entire field of scientific dream research. This shifts the burden of proof: while reports of dreamless sleep experience may seem to be an easy target, if only because of the novelty and alleged remoteness of Thompson's proposal for investigating dreamless sleep experience, we can now see that the proponent of the default view

will in fact have to take on the entire field of (report-based) scientific dream research as well. This raises the bar considerably; but first, more has to be said about how the methodological background assumptions of scientific dream research actually parallel questions asked in the classical Indian debate.

To begin with, note that the transparency assumption is analogous to the Advaitin and Yoga claim that upon awakening from dreamless sleep, we can veridically remember and report that we experienced nothing during sleep. To be sure, this type of report describes an experience marked by the absence of the complex imagery and narrative contents that characterize dreaming. Yet, in the Advaitin view, these are reports of an experiential state: in reporting having slept dreamlessly, we are reporting that we *experienced* nothing, in the relevant sense, during sleep;¹¹ we are not reporting the absence of experience. Thompson suggests that in order to turn the Advaitin view into a research strategy, the most reasonable and cautious approach is to assume that dreamless experience exists only intermittently, rather than persisting throughout dreamless sleep. The frequency with which dreamless sleep experience is reported to occur upon awakening will then be regarded as indicative of the actual occurrence of such experience. This is analogous to the reportability assumption. To endorse the stronger claim that dreamless sleep experience persists throughout sleep, at least prior to empirical investigation, would be to legislate an answer to the question of dreamless sleep experience. The weaker claim complements the assumption, implicit in scientific dream research, that periods of dreaming contrast with periods of nondreaming, which is quite different from saying that dreaming persists throughout sleep.

By combining my analysis of the methodological background assumptions of scientific dream research with Thompson's proposal on

the investigation of dreamless sleep experience, we can see that if we were to translate the Yoga and Advaitin view into a research methodology, we would find it to rely on assumptions that run parallel to those of scientific dream research. Dreamless sleep experiences, or so a modern-day, scientifically-minded Advaitin would be forced to admit, are reportable experiences; and if it should happen that (under sufficiently ideal reporting conditions, such as immediately after having awakened from sleep) one were unable to recall any such experience having happened during sleep, this would indicate that no such experience had occurred.

This also tells us that reports of non-dreaming should be further qualified: reporting the absence of experience during sleep is not the same as reporting dreamless sleep experience. The former is an instance of reporting an absence of experience, the latter is an instance of reporting a form of experience characterized by the absence of intentional objects; but it is still an experience report. Yet, while this requires terminological adjustments and shows that the concept of reporting a state of nondreaming is ambiguous, this adjustment is consistent with the familiar methodology; indeed, it falls out of the methods already used in dream research, when they are applied to the target of dreamless sleep experience.

From this, we can conclude that the default view of dreamless sleep as being characterized by the absence of subjective experience is intrinsically flawed for two related reasons. The first is that by treating dreamless sleep experience as a conceptual absurdity rather than as an open and empirically tractable question, it misconstrues the nature of the question of dreamless sleep experience. The second is that it stands in outright contradiction to the assumptions implicit in the scientific investigation of conscious experience during sleep. Dream research, understood as the scientific investigation of conscious experience during sleep, should be expanded to include dreamless sleep experience as well. And while this certainly will involve an adjustment of its conceptual resources, the good news is that its existing methodological background assumptions can remain largely intact.

¹¹ At this point, it might be objected that this formulation rides on a reification of the word "nothing", as if "nothing" itself could be turned into an object of experience. I return to this problem in section 4; as will hopefully become clear, my own positive model of dreamless sleep experience avoids this problem by introducing a qualified reading of what is described, in the Advaitin view, as experiencing or knowing nothing.

3.2 The Indian debate revisited: Lessons for the philosophical debate on the trustworthiness of dream reports

The analogy between the Indian debate on dreamless sleep experience and the background assumptions of scientific dream research not only highlights the inconsistency of the default view. There are also valuable lessons to be learned in the other direction, and considering the historical Indian debate can enrich contemporary debates on the status of dream reports as well. In particular, note that it is one thing to say that scientific dream research is implicitly committed to the transparency and reportability assumptions; but it is another to say that these assumptions are also theoretically justified. Elsewhere, I have defended the view that explanatory considerations justify the transparency and reportability assumptions: construing dream reports as (largely veridical) memory reports provides a *better explanation* of dream reporting behavior than skeptical alternatives that construe dream reports as the result of inference, misremembering or outright confabulation (Windt 2013, 2015, chap. 4). Here, I want only to point out that similar considerations apply to reports of dreamless sleep experience. In fact, Thompson's response to the Nyāya argument against dreamless sleep experience shows that casting reports of having slept dreamlessly as based on inference rather than memory is not a proper explanation at all. Instead, it leads to an argument that either results in an infinite regress or is circular. Again, there is a striking similarity to a similarly skeptical account of dream reporting from the 20th century. This time, the analogy with the historical Indian position will lend additional support to anti-skepticism about dream reporting.

To see why, another brief excursion into the history of theorizing about scientific dream research is instructive. Let us consider Norman Malcolm's (1956, 1959a) skeptical argument against the claims that dreams are conscious experiences occurring during sleep and that dream reports transparently show this to be the case. This argument was a direct reaction to early attempts, following the discovery of REM sleep,

to operationalize dreaming as a REM sleep phenomenon. Malcolm's argument hinges on the conceptual claim that "if a person is in *any* state of consciousness it logically follows that he is not sound asleep" (Malcolm 1956, p. 21). According to Malcolm, even though we use the same language to describe dreams and waking experiences, dreams (or at least such dreams as occur during sound sleep, which Malcolm, again for conceptual reasons, takes to be representative of dreaming proper) are not experiences, and for the same reason dream thoughts, feelings, and emotions are not real instances of their kind. As Malcolm puts it,

if a man had certain thoughts and feelings in a dream it no more follows that he had those thoughts and feelings while asleep, than it follows from his having climbed a mountain in a dream that he climbed a mountain while asleep. (Malcolm 1959a, pp. 51-52)

Malcolm's view is complex and a detailed discussion is beyond the scope of this commentary; suffice it to say that one of its more controversial upshots is that dream recall is not a real instance of remembering experience during sleep. Instead, "statements of the form 'I dreamt so and so' are always inferential in nature" (Malcolm 1959a, p. 65): one infers that one has dreamt when one realizes, upon awakening, that the events one seems to remember did not in fact occur. This claim struck many of his critics as contradicting both the common-sense understanding and the phenomenology of dream recall (see Dunlop 1977 for a collection of some of the most important critical essays; see Windt 2013, 2015, chap. 1 for discussion). Elsewhere, (Malcolm 1959b) explains that he takes dream recall to be inferential not in the psychological sense of actually drawing this inference when we notice that we have dreamt, but in the sense that we could give grounds for our belief that we dreamt if pressed to do so. However, because he fails to clarify what exactly these grounds are, his account remains sketchy. By applying Thompson's reconstruction of the Nyāya syllogism to Malcolm's claim, it quickly becomes

clear that even a more complete reconstruction of the inference would be intrinsically flawed. The result would be something like this:

1. While I was sound asleep, I had no experiences, including sensations, conscious thoughts, feelings, beliefs, or emotions.
2. This is because (i) I was in a special state (that is, not awake) or (ii) I lacked the necessary means for having experiences, including sensations, conscious thoughts, feelings, beliefs, or emotions (that is, my senses and mental faculties were shut down).
3. Whenever (i) I am in a special state (that is, whenever I am not awake) or (ii) I lack the necessary means for having experiences, including sensations, conscious thoughts, feelings, beliefs, or emotions (whenever my senses and mental faculties are shut down), I do not have experiences, including sensations, conscious thoughts, feelings, beliefs, or emotions.
4. As in the case of fainting or a blow to the head.
5. While I was sound asleep, I had no experiences, including sensations, conscious thoughts, feelings, beliefs, or emotions.

Malcolm concludes that sound sleep is comparable to other states of unconsciousness, and “to a person who is sound asleep, ‘dead to the world,’ things cannot even seem” (Malcolm 1956, p. 26).

If we follow this reasoning, then dream reports cannot ever be veridical experience reports: if we cannot have thoughts, feelings or emotions during sleep, then we also cannot have them during dreams, and we cannot actually remember (or veridically report) having had them after awakening. Rather, we sometimes awaken with the impression of having had such thoughts, feelings and emotions during sleep; and when we realize that they did not in fact occur, we infer that we dreamt.

To be fair, there might well be cases in which dream recall does have such an inferential nature. To use Malcolm’s example, it seems possible that I could awaken with the particularly vivid impression of having climbed a mountain

and then might realize, from the simple fact that I was lying in bed and nowhere near a mountain, that I had not actually climbed a mountain, but had been asleep. However, even if I was now quite sure that I had merely dreamt that I had climbed a mountain, it would not follow that the thoughts and feelings I remember having in the dream did not really occur. In order to draw this further inference, I would have to know that dreaming is a special state that is devoid of any experiences whatsoever.¹² As is the case for the Nyāya syllogism, this immediately invites the dual threats of circularity and of infinite regress: If I say I was in a special state because the thoughts and feelings I experienced in my dream were not real instances of their kind, I am reasoning in a circle. And if I say that I was in a special state because the mental faculties required for having thoughts and feelings were shut down (or because, as would better befit Malcolm’s argument, I had temporarily lost the capacity for producing the types of behavioral evidence that would enable another person to verify that I had been dreaming), then independent evidence would be needed—and so on. Again, without appealing to memory, no such evidence is available.

At this point it might seem that there is an easy solution: Perhaps, independent evidence for saying that dreaming is a special state has, in the meantime, become available. Malcolm’s analysis of dreaming was a direct reaction to early studies, discussed in section 3.1, on the correlation between REM sleep and dreaming, and his argument made much of the alleged impossibility of acquiring independent evidence,

¹² Incidentally, note that if it were the case that dreams are devoid of any experiences whatsoever, it would be utterly mysterious why we should awake with the vivid impression of having had such experiences in the first place. Indeed, Malcolm provides no explanation of why this happens. By contrast, my erroneous impression of having climbed a mountain during sleep is nicely explained by saying that during sleep, I had experiences that were sufficiently similar to their waking counterparts to create this impression. Again, this comes back to the idea that explanatory considerations favor the view that dream reports are actual memory reports, and not inferential. Perhaps, the difference between dreams that are belief-inviting beyond the borders of sleep, for instance by making us actually believe, if only for a moment, that the corresponding events actually occurred, and more commonplace dreams that do not induce such false beliefs can even be described in phenomenological terms (for a first proposal of how this might be done, see Windt 2015, chap. 10).

over and above dream reports, for the occurrence of dreams during sleep. Among Malcolm's critics, there was widespread agreement that he was simply mistaken about this latter point: sleep behavior, (for instance in patients with REM sleep behavior disorder, who are thought to act out their dreams due to a loss of REM sleep-related muscular atonia; see [Schenck 2005](#), [Valli et al. 2012](#)) sleep talking, and also polysomnographic measurements were (and continue to be) thought to provide exactly such independent evidence, perhaps even to the point of enabling researchers to verify dream reports (see for instance [Ayer 1960](#); [Rosen 2013](#); signal-verified lucid dreams are another example, as proposed by [Revonsuo 2006](#); see sec. 5.1 for a fuller discussion). Yet, even though the appeal to scientific dream research slightly changes the content of the argument, this merely restates the familiar syllogism, including its problems in a new guise.

To see why, let's say that rapid eye movements had indeed been found (as stated by the so-called scanning hypothesis; see [Dement & Kleitman 1957](#), to be directly related to visual dream imagery. Could we now analyze these eye movement patterns to diagnose the occurrence (and perhaps even the content) of dreaming even in the absence of (or in contradiction to) dream reports (see [Dennett 1976](#) for the discussion of this possibility)? Note that this is not an abstract philosophical issue: dream researchers have long dreamt the dream of moving beyond dream reports in the study of dreaming altogether. This ranges from science fictional visions of televising dreams ([Hall & Van de Castle 1966](#)) or of perhaps modeling them as an immersive, interactive virtual environment, as in [Antti Revonsuo's](#) (2006, pp. 300-303) dream catcher test, to real-world attempts to predict the content of dream reports from behavioral ([Leclair-Visonneau et al. 2010](#)) or neuroimaging ([Horikawa et al. 2013](#)) data. Again, the idea is that in the future, the analysis of neuroimaging data might be a way to verify dream reports, or even to move beyond their collection and analysis altogether. Elsewhere, I have argued that such attempts are circular: Dream reports, under the assumption of transparency, are used to

identify potential sleep-stage and neural correlates of dreaming; but the evidence such potential correlates provide is only as strong as the correlation, and so one cannot then turn around and use such measures as independent evidence to verify dream reports. Now, the Nyāya syllogism and its failure present a nice and crisp illustration of why this is the case. I think this is a nice example of the fruitfulness of a cross-cultural perspective on the methodological and conceptual issues involved in studying the occurrence of consciousness during sleep.

But there is another lesson to be learned. This is that the Nyāya syllogism is not an outdated problem, but one that persists even if we place it in the context of scientific dream research. The question of whether reports of having slept dreamlessly are experience reports or inferential is not of mere theoretical interest, but makes a real difference: assuming such reports, at least when given under ideal reporting conditions, to be veridical memory reports is the condition for a report-based scientific investigation of the relevant experiences in the first place. The historical debate, and [Thompson's](#) reconstruction of it, nicely highlights the need for acknowledging the relevance of first-person reports. Together, they also strengthen the theoretical case against skepticism about the trustworthiness of dream reports. With this anti-skeptical account in place, we can now move forward. In the next section, I sketch the outlines of a conceptual framework for describing dreamless sleep experience and its relation to dreaming.

4 From minimal phenomenal selfhood to minimal phenomenal experience: Towards a conceptual model of experience during dreamless sleep as pure subjective temporality

If what I have said so far is on the right track, then the question of whether dreamless sleep, at least on occasion, involves phenomenal experience is open to empirical investigation, and progress towards answering it can be made by applying the methods already used in scientific dream research, for instance by combining

timed awakenings in the sleep laboratory with questionnaires that are carefully calibrated to direct participants' attention towards the relevant features of such experiences and facilitate their reportability. Even occasional reports of dreamless sleep experience will support the claim that dreamless sleep experience exists. The next step towards turning the question of dreamless sleep experience into a scientifically tractable research project is to draw a more precise conceptual map of the territory. Sketching at least the rough outlines of such a conceptual map is my aim in this section.

Thompson's reconstruction of the classical Indian debate as well as his own positive proposals for how to study dreamless sleep experience provide a helpful point of departure. To begin with, as [Thompson](#) points out, the concept of dreamless sleep itself requires phenomenological refinement (p. 13). If dreamless sleep experience exists, then it is not enough to characterize dreamless sleep by the absence of dreaming or its electrophysiological correlates. Rather, dreamless sleep can now be seen to be a blanket term covering different types of conscious and nonconscious mental activity. Some forms of conscious mental activity that are commonly contrasted with dreaming (and in this simple sense can be said to occur in dreamless sleep), such as hypnagogic imagery during sleep onset or repetitive and non-progressive types of sleep thinking, are not candidates for the kind of dreamless sleep experience described in the Indian debate. Dreamless sleep experience in this narrow sense, if it exists, is a form of phenomenal experience characterized by nonintentional awareness ([Thompson 2015](#), p. 2): "When we're deeply asleep [...] we don't cognize anything—there's no object being cognized and no awareness of the 'I' as knower. Nevertheless, [...] we feel this absence while we sleep and remember it upon awakening" ([Thompson 2015](#), p. 238). Dreamless sleep experience is not just characterized by the absence of certain object-directed forms of conscious experience, but by the fact that this is an *experienced* absence. Moreover, it is not just the objects of experience that are absent, but also the subject of experience, or the "I". A very basic experiential fea-

ture, namely that of being an epistemic agent or a potential possessor of knowledge, has been lost (cf. [Metzinger 2013](#) for a fuller discussion of the term of an "epistemic agent model"). Dreamless sleep experience is characterized by a dissolution of subject-object duality, or, to put a more contemporary gloss on this, by a breakdown of even the most basic form of the self-other distinction ([Windt et al. 2014](#)).

This last point is important because it suggests a way of differentiating between dreaming and dreamless sleep experience. Many different definitions of dreaming exist—indeed, the lack of a uniform definition is an important desideratum for theoretical and experimental work on dreaming—but work on dreaming in philosophy of mind often focusses on a structural feature of dream experience. The assumption that dreaming involves the experience of a self in a world marks a point of convergence for philosophers of different stripes, ranging from contemporary philosophers of mind working towards an empirically informed theory of dreaming ([Metzinger 2003](#); [Revonsuo 2006](#)) to authors working in the tradition of classical phenomenology ([Husserl 2006](#); [Conrad 1968](#)).¹³ Studies have shown that an overwhelming majority of dream reports describe the presence of a dream self ([Strauch & Meier 1996](#)) though the precise way in which the dream self is represented is variable ([Occhionero et al. 2005](#); [McNamara et al. 2007](#)). The description of dreams as involving not just a self in a world, but an intersubjective world has even informed theories on the functions of dreaming (see for instance

¹³ Note that this way of thinking about phenomenal selfhood is quite different from the way the term "self" is used in the classical Indian literature. In his reconstruction of the Advaita Vedānta concept of witnessing, [Fasching \(2010\)](#) notes that the "witness" (sāksin) is not understood as an observing entity standing opposed to what it observes, but as the very taking place of 'witnessing' itself, and 'witnessing' is nothing other than the taking place of the experiential *presence* of the experiences, in which the experiences have their very being-experienced and thereby their existence." (p. 204) In this conception, "the 'self' is nothing other than becoming aware of experiential presence (consciousness) as such" (p. 207); it is "not a structural moment of what is given, but is the *very taking place of givenness itself*" (p. 210). Recall that one of the points of agreement between the Advaitins and the Nyāyas was that the self persists throughout sleep. But this is not the reading of the concept of "self" that [Thompson \(2014](#); see for instance chap. 10) has in mind when he says that in dreamless sleep experience, there is no longer an awareness of the "I", or what I mean when I speak of phenomenal selfhood.

Revonsuo et al.'s 2015 theory of dreaming as a simulation of social reality). Importantly, the description of dreaming as the experience of a self in a world also informs Thompson's own work on dreaming. In *Waking, Dreaming, Being*, he tells us that "the core feature of full-blown dreaming is the experience of immersion in the dream world" (Thompson 2014, p. 127), and also that this immersive quality is exactly what distinguishes hypnagogic imagery during sleep onset from dreaming (pp. 135ff.). The hypnagogic state is a state of absorption, in which attention is fully captured by a series of dynamically changing and often short-lived images; "the hypnagogic state blurs the boundaries between inside and outside, self and world" (p. 124).

This description coincides nicely with my own theoretical work on dreaming. Elsewhere, I have argued that the analysis of self-experience is the key towards understanding not just different types of dreaming (Windt 2010, 2015, chap.s 11 and 12), but also the relationship between dreaming and waking experience. In this view, the common denominator underlying different types of dreams, such as lucid and nonlucid dreams, but also nightmares and false awakenings is their immersive quality. Even in simple forms of dreaming, there is still a sense of presence, a phenomenal *here*, or the sense of being located at a specific point in space, as well as a sense of duration centered on a phenomenal *now*. This basic structure is preserved even when the features that characterize a majority of dreams, such as interaction with non-self dream characters, objects, emotions, or even visual imagery are lost. In such minimal dreams, phenomenal selfhood takes the form of pure spatiotemporal self-location, arising independently of more complex forms of phenomenal selfhood that involve the experience of being a thinking self and embodied agent. There may even be the experience of phenomenal disembodiment, or of lacking a body, and the dream self may be experienced (and later described) as an abstract, undefined volume of indeterminate extension or even as an unextended point in space. Even though this sense of identification with a phenomenal *here* and *now* in-

volves a drastically reduced form of phenomenal selfhood, it is still sufficient to ground retrospective claims of having had a self in dream reports. The basic structural feature of a self that is experienced as distinct from and located at a precise point within the world is preserved. To be sure, the locus of self-location and self-identification is more fluid in dreams than in wakefulness—the phenomenal *here* is subject to sudden shifts, and sometimes, we identify with a dream character or even a series of dream characters that are quite distinct from our waking self (Rosen & Sutton 2013). Yet, as long as there still is a world experienced as distinct from the self, at least a basic form of the self-other distinction continues to exist.

Within this framework, immersive spatiotemporal hallucination, or self-location with respect to a largely nonveridical, spatiotemporal reference frame, marks the cutoff line between dreaming and nondreaming. It also helps isolate and empirically ground minimal phenomenal selfhood (Blanke & Metzinger 2009), or the simplest conditions under which the experience of being or having a self arises. Here, I would like to suggest that this framework can be extended to dreamless sleep experience as well. A very basic point is that we can now sharpen the claim that dreamless sleep experience is a selfless state. Within the present framework, in order for dreamless sleep experience to count as selfless, even the basic form of self-other distinction that underlies spatiotemporal self-location must be lost. The next step is to consider the spatial and the temporal characteristics of self-location independently of each other and ask whether either of them, considered on their own, would be sufficient to give rise to phenomenal selfhood. An affirmative answer would mean that we had not yet identified the phenomenal signature of dreamless sleep experience; an even more simplified account would be needed.

Considering the spatial and temporal aspects of self-location separately, there seems to be a strong conceptual link between the phenomenal *here* and the sense of being located in and relative to a larger spatial expanse. A spatial reference frame, according to the present

theory, turns into an experienced world when it is centered on a phenomenal *here*, which in turn is identified as the self. The spatial variant of presence thus seems to have the self, or some rudimentary form of self-other distinction, written into it. Moreover, the attempt to conceive of an experience characterized by a phenomenal *here* but lacking any temporal characteristics whatsoever strains the limits of conceivability. Speaking of an experience that is both instantaneous, lacking any temporal extension, and fails to have temporal location seems to be a contradiction. It is not clear how this could count as an experience at all, and even less how it could count as a reportable one.

By contrast, the phenomenal *now* does not appear to carry the same conceptual commitments. At least intuitively, the notion of a form of temporal experience that is independent of and perhaps more basic than the experience of being or having a self seems more acceptable than that of an immersive but nonetheless selfless form of spatial experience. Moreover, we can at least conceive, it would seem, of a phenomenal *now* that fails to be differentiated from or clearly located relative to a larger temporal reference frame.¹⁴ And we can also, it would seem, conceive of an experience characterized only by temporal but not by spatial characteristics. Thinking, for instance, is not always experienced as having spatial location (as in thoughts occurring in one's head), but it certainly has temporal dynamics.¹⁵ Spatiality does not seem to be essential to phenomenality in quite the same way as temporality.

Note that I do not intend these conceptual considerations to carry too much weight. In the framework I am working towards, conceptual

distinctions are informed by differences in the structure of phenomenal experience and such differences should at least in principle be memorable and describable, for instance in dream reports or reports of dreamless sleep experience.¹⁶ I also think that the most empirically plausible view will allow for gradual transitions between states involving a phenomenal self and those retrospectively described as selfless; and the same may also be true for the emergence of the simplest forms of phenomenal experience. If this is correct, then we should expect there to be a certain amount of uncertainty when dealing with borderline cases. Where exactly to draw the cutoff line for minimal phenomenal selfhood in a given case may well be hard (if not impossible) to determine; but even so, it might still be useful to introduce a conventional cutoff line (for instance by saying that minimal phenomenal selfhood involves both the spatial and the temporal aspects of self-location) if this helps pick out a theoretically meaningful transition in the structural features of experience and guides future research in a constructive manner. We will also expect such a theoretical conception to be reasonably well aligned with the way such experiences are described in retrospective reports.¹⁷ I think that both types of considerations support the claim that spatiotemporal self-location can be meaningfully described as a minimal form of phenomenal selfhood, or at least as a theoretically salient point of transition on the trajectory from states described as

¹⁶ This is not to deny that experiences (or qualitative aspects of experiences) could exist that are beneath the cutoff line of memorability and reportability. Certain subtle aspects of phenomenal experience, such as hues of color, do seem to outrun our ability to categorize and reidentify them over time (Raffman 1995). Here, I am only claiming that such subtle aspects of experience are not candidates for the report-based type of scientific investigation I am interested in here.

¹⁷ This is a prediction, and different subjects may mean different things when they describe an experience as selfless. For some this may mean an experience characterized by spatiotemporal self-location, but in which they had the experience of being a disembodied entity (cf. Windt 2015, chap. 7); others may describe episodes characterized only by their temporal features as involving a self. There is also the familiar problem that reports of selfless experiences easily slip into a performative self-contradiction, of the type "I had a dream in which I was not present"; such episodes are clearly remembered and reported by someone. But we should not expect the folk-psychological use of terms such as "I" or "self" to align perfectly with a particular technical definition. This is a good example of where specific interview questions might increase the expressive granularity of retrospective reports.

¹⁴ In fact, if we conceive of temporal experience as involving a specious present, we might say that the phenomenal *now* simply is identical with a rudimentary form of a temporal reference frame. I return to this point later. Alternatively, if we conceive of temporal experience as consisting of a series of unconnected moments that themselves have no temporal extension, then again it would seem that each of these could occur in isolation and without being embedded in a larger temporal reference frame.

¹⁵ This phenomenological observation is reflected in the classical idea that the mind cannot be spatially located in the physical world. Mental states persist over time, but they do not have spatial characteristics such as expansion or separable parts. Perhaps, this phenomenological observation lies at the root of metaphysical claims about the relationship between mind and body.

selfless to states involving self-experience in a fuller sense. By contrast, the phenomenal *now*, when it arises independently of spatial self-location, is a candidate for a structural feature of phenomenal experience that provides the conditions of possibility for self-experience but that when occurring on its own is still prior to it. I would like to suggest, then, that pure subjective temporality is a candidate for minimal phenomenal experience; it is a condition for but still more basic than minimal phenomenal selfhood. It can be described as subjective only because it involves phenomenal experience; yet, it does not involve the additional experience of being a self, or a separate entity having the experience.¹⁸

There is, of course, a rich philosophical debate on the nature of time experience, as well as a large empirical discussion (for an introduction, see [Dainton 2010](#); [Arstila 2014](#); [LePoidevin 2015](#)). I cannot begin to do justice to this literature here, but want only to focus on one specific aspect. This is the idea, which we find in William James as well as in Husserlian phenomenology, but also in the neuroscience of time consciousness (see for instance [Pöppel 2003](#)), that even the smallest unit of temporal experience, the temporal *now*, is extended rather than instantaneous.¹⁹ Following this con-

ception, a rudimentary form of duration would be intrinsic to the phenomenal *now*; and neuroscientific work seems to suggest that this temporal *now* is itself variable ([Wykowska & Arstila 2014](#)). The window of simultaneity, or the maximum time-frame within which two different events are experienced as occurring *now*, is modality-specific. The cutoff line for two stimuli being experienced as simultaneous is, for instance, larger for visual stimuli than for auditory ones. As [Wykowska & Arstila \(2014, p. 443\)](#) note,

it might be that a relatively broad window of simultaneity is actually beneficial. The human brain needs to exhibit some degree of tolerance to asynchronous stimuli in order to be able to bind different sensory inputs into one event. The window of simultaneity can be seen as an integration window for stimuli and, as such, is a necessary mechanism for binding signals from different pathways into one single object or event.

Human temporal resolution is flexible, it is easily affected by attentional processes as well as by training and expertise ([Wykowska & Arstila 2014](#)). Duration perception might be state-dependent as well, showing characteristic changes in altered states of consciousness and psychiatric disorders ([Noreika et al. 2014](#)); and perhaps the same is true for the degree to which the experienced *now* itself is stretched in time. There also seems to be a close relationship between changes in time perception and alterations in self-experience. When the self becomes the focus of attention, when we attend to our current mental or emotional state, or to bodily sensations (such as hunger or pain), time seems to slow down; by contrast, when we are thoroughly absorbed in an activity, time contracts and seems to move faster ([Wittmann in press](#)). When self-experience is lost, as in selfless states, the loss of a reference point may be associated with feelings of timelessness ([Wittmann in press](#)); the phenomenal *now* is stretched indefinitely. There is a sense of duration, but the sense of

¹⁸ Note that this is related to a terminological difficulty that is implicit in the Indian debate, as well as in Thompson's reconstruction of it. As noted earlier, both sides in the Indian debate assumed the self to persist throughout sleep; they merely disagreed whether the self is necessarily conscious. My proposal that we redescribe dreamless sleep experience in terms of pure subjective temporality captures this idea that the self persists in a thin sense even when awareness of any intentional contents is lost. At the same time, recall that dreamless sleep experience is thought to be characterized by a collapse of subject-object duality and by an absence of any intentional objects of awareness. In this state, nothing, including the self, is thought to be known or cognized. There is no longer an individual, consciously experienced first-person perspective. It is this thicker and more substantial notion of a self experienced as distinct from other objects or persons that I propose is lost in dreamless sleep experience; the persistence of such a self would mean that there would still be an intentional object of awareness, and thus would indicate a more complex state than that characterized in the Indian debate as dreamless sleep experience.

¹⁹ This could, of course, turn out to be false. Even if the underlying neural representations are temporally extended, the same may not be true of conscious states themselves; these may still be conceived of as elementary and momentary events lacking spatial or temporal structure. For a recent defense of such a view, inspired by the Abhidharma doctrine of momentariness, see [Chadha \(forthcoming\)](#). Yet, even if the experience of continuity and persistence over time turned out to be an illusion, this would still be an interesting structural feature of phenomenal experience. For present purposes, the basic phenomenological claim, according to which the phenomenal *now* is temporally stretched rather than momentary and discrete, is enough.

succession, of there being a chain of present moments, has been lost.

Importantly, this way of thinking about subjective temporality and the experienced *now* is one which Thompson (2015) endorses. He explicitly appeals to the Husserlian conception of time experience in his defense of retrospective reports of dreamless sleep experience. Here, he suggests that memories of dreamless sleep experience may be grounded by retentional awareness, “the holding onto the just-past as an intentional content belonging to our consciousness of the passage of time, including our mental lives as flowing in time” (p. 9). Because temporal experience has the retention of the immediate past and protention, or the anticipation of the next moment, written into it, the moment after awakening still carries with it the traces of dreamless sleep experience: “Immediately, the ego sense appropriates the lingering impression or retention of not-knowing and associates this retention with itself, thereby generating the retrospective thought, ‘I did not know anything’” (p. 10).

In *Mind in Life*, Thompson (2010) endorses a version of Husserl’s conception of time-consciousness according to which the streaming, flowing character of subjective experience is both the “condition of possibility for every other kind of consciousness, but is not itself made possible by some other, still deeper level of consciousness” (p. 324). This absolute flow of consciousness is self-constituting (p. 324); it is also prior to and essential for phenomenal selfhood. As Thompson (2010) puts it,

to be aware of phenomena across time, consciousness must be retentionally and protentionally aware of itself across time. Therefore, time-consciousness entails prereflective self-awareness. In other words, our being conscious of external temporal phenomena entails that our temporally enduring experiences of those phenomena are self-aware. Inner time-consciousness is thus nothing other than prereflective self-awareness. (p. 328)

This prereflective awareness that consciousness has of itself (its self-luminousness, reflexivity, or

self-acquaintance²⁰) is not yet the same as being or having a phenomenal self in the sense used here. Rather, this minimal form of phenomenal experience is the condition for the emergence of minimal phenomenal selfhood.

My suggestion, then, is that we can enrich our theoretical conception of dreamless sleep experience by applying Thompson’s account of how we *remember* dreamless sleep experience (namely with the help of retentional awareness) to the description of dreamless sleep experience itself. Dreamless sleep experience involves pure subjective temporality that is not yet structured around intentional objects, including a phenomenal self. As Thompson (2015) puts it, “although deep sleep creates a gap or a rupture in our consciousness, we often feel the gap immediately upon awakening. [...] We are aware of the gap from within our consciousness” (p. 4). Just as upon awakening, I am directly aware that it was I who was asleep and unknowing, I am typically aware that a certain (though perhaps indefinite) amount of time has passed. Following Proust’s more poetic formulation in the passage quoted by Thompson,

a sleeping man holds in a circle around him the sequence of the hours, the order of the years and world. He consults them instinctively as he wakes and reads in them in a second the point on the earth he occupies, the time that has elapsed up to his

²⁰ Again, there are subtle terminological differences. For instance, Williford (2015a, pp. 10–11; see also Williford 2015b) writes that reflexivity or self-acquaintance is “an essential structural feature of all consciousness; and I take it to be a phenomenological datum. All streams of consciousness are immediately aware of themselves, and that is the foundation of all other forms of self-representation, autobiographical cognition, and so on. This reflexivity is subjective character (for-me-ness), but it is a mistake to turn this structural feature into a kind of entity or homunculus.”

My account is compatible with much of what Williford says here; I agree that we are considering a basic and essential feature of conscious experience, and one that should not lead us to posit an independent entity that is identified as the self. Yet, I think there is room for phenomenal selfhood as a structural feature of experience over and above the reflexivity of even the simplest kinds of phenomenal states. Even readers who disagree with my description of this as a form of phenomenal selfhood might still agree that the target property of spatiotemporal self-location is distinct from the more basic reflexivity of consciousness. Adopting the conceptual convention of describing this as a form of self-experience does not, I take it, require us to reify the self or to slip into a homuncular view, but simply offers a conceptual tool for describing the way we experience ourselves as being or having a self.

waking; but their ranks can be mixed up, broken. (p. 3)²¹

We might even say that metaphorically speaking, subjective temporality provides a reference frame that is still empty, but poised to integrate and lend temporal structure to intentional contents such as thoughts, objects and events, but also the self, as they arise—for instance by imposing sequential order on them and representing some of them as simultaneous, and others as successive. Yet, this form of temporality is more basic than the events it later integrates; it predates them and provides a space in which they can appear.

Incidentally, this idea fits in nicely with the Vedantan view that, “deep sleep is a kind of ‘ground state’ of consciousness, a lowest-energy state from which the ‘excited states’ of dreaming and waking arise” (Thompson 2014, pp. 260-261). Again, deep sleep is the baseline, the causal source from which other conscious states arise; it is also called “seed sleep”, because it is thought to contain the seeds of both dreaming and waking consciousness. Perhaps we can begin to make sense of this idea by saying that dreamless sleep experience, understood as pure subjective temporality, is a candidate for minimal phenomenal experience.²²

²¹ A prediction that seems implicit in Proust’s observation that if we are suddenly overcome by sleep, we no longer know what time it is upon awakening, is that dreamless sleep experience may bear an interesting relation to the ability to estimate how long one has slept. Perhaps, intermittent periods of dreamless sleep experience even ground our awareness that some time has passed or are responsible for the ability, which may be more pronounced in certain subjects, to awaken just before the alarm clock goes off (thanks to Valdas Noreika for pointing this out). By contrast, if we awaken from a state lacking any form of phenomenal experience whatsoever—as in some forms of anesthesia—there may be no sense of a preceding temporal gap and a more profound sense of temporal disorientation. At present, this is, of course, entirely speculative, but it might be a question worth asking.

²² The temporally dynamic nature of experience is also of central importance for understanding the neural correlates of conscious experience. As Melloni (2015) points out, while the mechanisms for updating the contents of consciousness have been investigated by numerous studies, the mechanisms underlying the maintenance or the flow of conscious experience fall outside the scope of most existing paradigms. She also proposes that the temporal flow of consciousness is a fundamental property of experience and an important target—perhaps the most important target—for future research on the neural correlates of consciousness. Similarly, Noreika (2015) suggests that focusing on the analysis of individual contents of consciousness, as is standardly done in mainstream research on the neural correlates of consciousness, misses the temporality of consciousness; instead, to make progress toward understanding this more fundamental property, he proposes contrasting conscious and nonconsciousness states.

How can we make progress on identifying real-world cases of dreamless sleep experience? Importantly, if the account of dreamless sleep experience defended here is even remotely correct, we should not expect dreamless sleep experience to be restricted to experienced meditators. Instead, dreamless sleep experience might be fairly prevalent even in people without any formal training in contemplative traditions. This approach requires disambiguating between at least two variants of the target phenomenon. Note that within the Indian conception of dreamless sleep, we can distinguish between an insight component and a more basic experiential component. The insight component refers to the ability to become aware, during sleep, of the nature of this state. This is not necessarily a conceptually mediated form of knowing *that* you are currently sleeping dreamlessly, but rather consists “in being able to witness the state of dreamless sleep and recall its phenomenal clarity upon awakening” (Thompson 2015, p. 15). Still, even this nonconceptual form of witnessing is not epistemically neutral, but can lead (or fail to lead) to veridical retrospective reports. To be sure, this form of insight or awareness itself can have a particular phenomenal feel—it bears the phenomenal signature of knowing (Metzinger & Windt 2014, 2015), the feeling of just having become aware of the nature of one’s ongoing state—but importantly, this type of phenomenal experience carries with it epistemic commitments. My feeling of knowing can be true or false. It also seems plausible, as suggested by Thompson, that meditation facilitates this type of lucid dreamless state, or perhaps could even be a way of inducing it systematically.

But the model of dreamless sleep experience as pure subjective temporality also points to a more basic experiential component that as such bears no obvious connection to an epistemic state of knowing or of being aware of the nature of the state one is currently experiencing. Dreamless sleep experience in this primary phenomenological reading refers to a kind of experience during sleep; but this does not require the ability to conceptualize this *as* a form of sleep experience. In principle, you can have dreamless sleep experience without realizing

that you are asleep: dreamless sleep experience is a form of experience occurring *in* sleep, but it is not necessarily an experience of sleep *as* a state of sleep. It might enable us to estimate how long we have slept, but it can also be misleading, maybe even leading us to misjudge whether we have slept at all. This is particularly obvious if dreamless sleep experience is construed as an answer to the question of how we know, upon awakening, that we slept peacefully (Thompson 2015, p. 4). Thompson's reconstruction of the Indian debate, taken together with my analysis, suggests that because this state is characterized only by its temporal character, we have the sense of there being a gap between two periods of wakefulness; and because this gap is devoid of intentional objects, we describe it as peaceful. Yet, this does not seem to require that we were aware of (or took ourselves to be aware of) the nature of this state while it was occurring, namely during sleep. If any sophisticated epistemological reading of insight were indeed crucial to dreamless sleep experience, the experience of having slept peacefully would have to be reserved for special subject groups, such as experienced meditators—and it would be quite mysterious why clearly, it is not.

Perhaps we can model the relationship between the epistemic and the phenomenological components of dreamless sleep experience on the relationship between lucid and nonlucid dreaming. Thompson (2015, p. 15) himself explicitly contrasts lucid dreaming, or knowing that you are dreaming while you are dreaming, with lucid dreamless sleep. Given this suggestion, a good place to begin the project of broadening the investigation of dreamless sleep experience beyond expert meditators is to consider reports from experienced lucid dreamers.

5 Candidates for pure subjective temporality during sleep

5.1 From lucid dreaming to lucid dreamless sleep?

Lucidity is commonly defined as awareness that one is dreaming while one is dreaming (for ex-

cellent reviews, see Voss & Hobson 2015; Dresler et al. *in press*). Often, this is associated with an ability to control not just one's own actions in the dream, but also the course of the dream, the actions of non-self dream characters, etc. In particular, lucid dreamers can signal that they have now become lucid by making prearranged patterns of eye movements, such as looking right – left – right – left within their dream. These gaze shifts correspond to the movements of their physical eyes and can be identified on the electrooculogram. This technique of signal-verified lucid dreaming enables researchers to identify the precise period of sleep during which certain actions were performed during a lucid dream and potentially to identify their electrophysiological and neural correlates (Dresler et al. 2011, 2012). Lucidity can occur spontaneously, but a number of methods for inducing lucidity are discussed in the literature (Stumbrys et al. 2012). There have even been suggestions and attempts to experimentally induce dream lucidity through electrical stimulation (Noreika et al. 2010a; Voss et al. 2014; Voss & Hobson 2015). While still in its early stages, this work clearly shows that lucidity is a robust phenomenon; and combined with the ability to control the dream as it unfolds, it makes laboratory studies of lucid dreaming compelling.

One reason for being interested in lucid dreams within the present context are reports of lucid dreams describing a loss of phenomenal embodiment, or even a dissolution of the self (see Windt 2015, chap.s 7, 11 for discussion). Some of these appear to fulfill the requirements for minimal phenomenal selfhood described earlier: in so-called imageless lucid dreams (Magallón 1991; Bogzaran 2003; Hurd 2008), self-identification may be relative to a disembodied point in space and can arise independently of bodily sensations and even of visual imagery (see also LaBerge & DeGracia 2000). While most of these reports, so far, are anecdotal, it is tempting to think that lucid dreams could be used to systematically investigate the transition from minimal phenomenal selfhood to more complex forms of self-experience involving the experience of being a thinking self and em-

bodied agent. Importantly, according to some of these reports, even this basic sense of self-identification and location within a larger spatial expanse can be lost. I would like to suggest that such cases may involve a shift from a simple form of lucid dreaming involving minimal phenomenal selfhood to lucid dreamless sleep experience. Here is a single example:

I am suspended in space—dream space, I think. There is nothing here, just millions of greyish dots and I am one of the dots, there's no dream-body anymore, I'm just a dot [of] pure consciousness suspended. A feeling of great peace comes over me and a sense of gentle, infinite expansion. It's as if everything and nothing are the same thing and there is a sense of effortless belonging. As the sense of expansion increases I am no longer a single dot of consciousness; all the dots are me and I am them. There's no "I" or "them." We are one. There's just a blissful sense of timelessness and oneness and a merging with the light. After an indefinable length of time, I start to feel the weight of my body in bed, and settle back into it, tingling all over. (Clare Johnson, unpublished dream report, March 19, 1995)

If we take the report at face value, it describes a gradual transition from minimal phenomenal selfhood, characterized by phenomenally disembodied spatiotemporal self-location, to selfless experience. This transition is accompanied by a sense of spatial expansion, in the course of which the sense of the self as distinct from the environment is lost. To the extent that there still is a sense of spatial self-location, this no longer involves the experience of being located relative to something else. There is also a change in the temporal structure of experience, almost as if the experiential present, the phenomenal *now*, had been stretched indefinitely. The period following the dissolution of the self is still experienced as having duration, but this duration is indefinable and no longer structured around any events.²³ Following Metzinger

(2013), we might want to describe this as involving a transition from a minimal unit of identification, in which an unextended point in space is described as the locus of the self, to a maximal unit of identification. In such cases of "pure consciousness", he suggests, the unit of identification is

the most general phenomenal property available for identification at all: Philosophers might call it the global "unity of consciousness", or phenomenality per se, or awareness as such, namely the singular, integrated, all-pervading quality of consciousness characterizing the current totality of experiential contents, as it is given in every single moment of experience. (Metzinger 2013, p. 5)

I would like to suggest that we can now be more precise. The moment at which self-location dissolves—or at which minimal phenomenal selfhood is replaced with the maximum unit of identification—involves a transition to the type of pure subjective temporality that earlier, I suggested might be the phenomenal mark of dreamless sleep experience. As lucid dreaming gives way to lucid dreamless sleep experience, minimal phenomenal selfhood shades into pure phenomenality, in which phenomenal experience is characterized only by its temporal structure. I find it telling that according to Johnson's report, this latter part of the episode appears to strain the limits of reportability, and also that despite its indefiniteness, the experience is described as blissful; again, this is exactly what the Indian focus on the experience of having slept peacefully would lead us to expect.

Clearly, this single dream report presents anecdotal evidence at best; still, I would like to suggest that a first step towards extending the investigation of dreamless sleep experience beyond experienced meditators might be to investigate imageless lucid dreams in experienced lucid dreamers. What makes me cautiously optimistic is that lucidity is often described as a very unstable phenomenon, as involving a balancing act

²³ A similar link between the dissolution of the self and the experience of timelessness, or of an indefinite duration, may exist in deep medit-

ative states (Berkovich-Ohana et al. 2013), but also, for instance, in drug-induced altered states of consciousness (Wittmann in press).

between maintaining lucid insight (rather than slipping back into a nonlucid dream or awakening) and remaining engaged enough in the ongoing dream to prevent it from dissolving completely (Brooks & Vogel song 2000). Lucid dreamers often describe that imagery can take on a faded, washed out quality, or that lucidity is followed by a period of darkness or, alternatively, of light; indeed, this may be why such reports often slip into mystical language to describe such experiences. Here, I want only to suggest that in such cases, the unwanted fading of lucid dream imagery may actually be an opportunity for experimentally investigating the transition to dreamless sleep experience.²⁴

Before moving on, I want to suggest that the comparison between lucid dreaming and lucid dreamless sleep is also interesting for another reason. This is that as is the case for nonlucid dreams, there continue to be a number of conceptual uncertainties about how to define lucid dreaming and whether to describe it as a genuine sleep phenomenon or as a hybrid state between REM sleep and wakefulness (Voss et al. 2009; for a discussion of lucidity and insight from a philosophical perspective, see Kühle 2015; see Voss 2015 for a critical reply). Also, while some authors consider any dream involving insight into the fact that one is now dreaming as lucid, others reserve the term lucidity for cases in which there is a marked increase in the overall vividness of multimodal imagery as well as a shift towards wake-like cognitive activity, including the ability to engage in rational thought, full recall of waking life, and insight into the fact that none of the events occurring within the dream have any real-world consequences (for a first attempt to test these different conceptions of lucidity experimentally, see Voss et al. 2013).

On the conception that I favor, lucidity is not necessarily accompanied by an all-pervading change in the phenomenal character of the dream; rather, lucid dreams are gradually distinguished from nonlucid ones along a number

of dimensions (Windt & Metzinger 2007; Nor-eika et al. 2010a; Voss et al. 2013). While laboratory studies, because of their reliance on signal-verified lucid dreams, necessarily focus on lucid dreams involving at least some form of control, the conceptually mediated insight into the fact that one is now dreaming is orthogonal to the other experiential qualities of dreaming. Insight is also necessary to score a given report as describing a lucid dream—but aside from this methodological fact, the ability to conceptualize one’s ongoing experience as a dream—to have the thought “I am now dreaming”—can coexist alongside the types of vivid, often bizarre and emotionally charged imagery and erratic reasoning that characterize a majority of nonlucid dreams as well. Lucidity can be the outcome of a conscious inference (of the type “this cannot be happening, so I must be dreaming”), but often appears to be driven by a sudden feeling, sometimes described by saying that the dream suddenly took on a dreamlike feel or a hyperreal character (see Windt 2015, chap. 9 for details and further references). Perhaps, this precursor to full, conceptually mediated lucidity is similar to the type of nonconceptual awareness that is thought to accompany lucid dreamless sleep experience as well. This suggests two further questions. The first is whether nonlucid forms of dreamless sleep experience exist as well. The second is whether in dreamless sleep experience, anything analogous to prelucid dreams exists. I discuss these in turn.

5.2 From white dreams to nonlucid dreamless sleep experience?

Again, we can approach the project of identifying candidates for nonlucid dreamless sleep experience by asking whether instances of minimal phenomenal selfhood exist in nonlucid dreams. If so, we could once more expect these to occur in the vicinity of minimal phenomenal experience during dreamless sleep.

A possible candidate for such a state are so-called white dream reports, in which subjects describe the impression of having experienced a dream but are unable to describe it in any detail. It seems plausible that a subgroup of white

²⁴ Similarly, in the tradition of dream and sleep yoga, dream lucidity is sometimes described as a preliminary stage of becoming aware of sleep; again, realizing that one is dreaming precedes the dissolution of dream imagery while maintaining awareness of dreamless sleep. See for instance Wangyal & Dahlby (1998).

dream reports can be explained by forgetting. Especially where the subject describes the distinct feeling of having had a complex dream but being unable to remember it in any detail, this would seem to be the most plausible interpretation. There is some reason for thinking, however, that this may not be the case for all reports of white dreaming. In at least some cases, the impression of having had some kind of experience prior to awakening, coupled with an inability to describe any particular aspects of the experience, such as any specific forms of imagery or narrative contents, might not be an artifact of forgetting, but might reflect the structure of the experience itself. At least a subset of white dreams might involve a sense of spatiotemporal self-location, or minimal phenomenal selfhood, arising in an otherwise imageless nonlucid dream. And if this were supported by future studies, then it might even make sense to ask whether perhaps, a further subgroup of white dreams could more properly be described as involving nonlucid dreamless sleep experience. In the current framework, these latter types of white dreams would not count as proper dreams at all: they would be instances of pure subjective temporality arising independently of the spatial aspects of self-location and self-identification. They would involve a form of minimal phenomenal experience that could no longer be described as minimal phenomenal selfhood, and thus as a dream. Perhaps, we occasionally really do retain some awareness, after awakening, of phenomenal experience having persisted during sleep. And perhaps, unable to remember any specific details, we then assimilate them to more familiar types of experiences, labeling them as white dreams.

Again, all of this is still extremely speculative and everything I have said so far about white dreams should be read, at best, as a careful prediction of what we might say in light of future findings. In particular, I do not mean to suggest that white dream reports, or a subgroup thereof, can already be regarded as examples of dreamless sleep experience: I only mean to propose that they are an initially promising target for future research on dreamless sleep experience. Still, these considerations fit in nicely with

the finding that white dreams are particularly frequent during slow-wave sleep. According to one study, awakenings from stages 2 and 3 NREM sleep were followed by roughly equal rates of dream reports, white dream reports, and reports of nondreaming (Noreika et al. 2009). Their occurrence in the vicinity of reports of dreaming and of nondreaming might indicate that white dream reports describe a transitional state between the two. Moreover, even dream reports obtained following awakenings from these sleep stages were often static, describing experiences lacking narrative progression as well as movement sensations (Noreika et al. 2009, 2010b). Participants sometimes described the sense of being present in a static scene, as in quietly sitting on a bench, with nothing else happening (Valdas Noreika, personal communication; see also Noreika 2014, p. 52). Even in the absence of narrative progression, there was still a sense of duration, and according to subjective estimates, these simple dreams lasted between thirty seconds and one minute.

An interesting possibility could be to investigate the wording of white dream reports in more detail. To my knowledge, this has not yet been done. Maybe there are indeed subtle differences in the wording of such reports, and perhaps these would enable researchers to distinguish cases in which there is the impression of having forgotten a complex dream from ones describing imageless and perhaps even selfless and objectless episodes of phenomenal experience. Again, it might be possible to increase the expressive granularity of reports with the help of training or specific questionnaires, thus rendering subtle phenomenological differences visible that would be otherwise overlooked. A possible finding could be that some of these experiences involve a continued sense of presence and self-location in an abstract, amodally experienced spatial expanse, whereas in others, even this basic sense of self is lost and only the feeling of duration, or of an indefinite temporal expanse, is present.

A particularly promising way to do this would be to use a serial awakenings paradigm, in which participants are awakened multiple

times throughout the night at intervals of 15-30 minutes, thus maximizing the number of reports that can be collected throughout the night (Noreika et al. 2009; Siclari et al. 2014; Siclari et al. 2013). Questions focusing on the temporal aspects of experience could then be used to identify those periods, if any, in which dreamless sleep experience is most likely to occur. For instance, Siclari et al. (2013) asked their participants to estimate how long they had been having continuous experiences before being awakened, but also how long their most recent experience had lasted, how far back in time they could recall any narrative events, and how rich and complex the experience was. They found that during stages N2 and N3, estimates for duration, recall back in time and richness were low. Still, these results could be influenced, in part, by the fact that the interview questions focused on the objects of consciousness and on narrative events. If the questions were reworded in such a way as to cover dreamless sleep experience, the patterns of responses might change. Even so, it is interesting to note that during sleep onset, there was a dissociation between these measures, with participants estimating a long duration of the last conscious experience, but a low richness and ability to recall back in time. At least, this suggests that the estimated duration of conscious sleep states does not always map cleanly onto the ability to recall specific contents. For now, I want only to suggest that a similar strategy could interestingly be applied to the investigation of dreamless sleep experience as well.

This is also attractive in view of the goals of this line of research. Note that Siclari et al. (2014) explicitly use the serial awakenings paradigm to contrast the presence and absence of conscious experience independently of task performance and within the same sleep stage (for a similar suggestion, see Noreika et al. 2009; Noreika 2015), the ultimate aim being to identify the task- and state-independent neural correlates of conscious experience. For this project, dreamless sleep experience, as a candidate for minimal phenomenal experience during sleep, is clearly a relevant target phenomenon.

5.3 From subjective insomnia to unwitting expertise of dreamless sleep experience?

The final example that I wish to discuss is sleep misperception in subjective (or paradoxical, as it is also sometimes called) insomnia. The term objective insomnia, reserved for patients suffering from actual sleep loss as conventionally measured, is sometimes contrasted with subjective insomnia, which refers to subjects who systematically underestimate the time they actually spend asleep (Harvey & Tang 2012; Perlis et al. 1997). This mismatch between subjective sleep perception and objective measures of sleep sometimes leads to a trivialization of subjective insomnia—and the suggestion that their diagnosis as insomniacs is somehow not “real” can be experienced as infuriating by those afflicted by it (Greene 2008). Subjective insomnia is clearly not an imaginary problem, but a cause of real suffering. In fact, patients with subjective insomnia may experience more severe impairments in cognitive functioning than insomnia patients who do not underestimate the amount of sleep they are getting. Furthermore, worrying about getting enough asleep may precede actual sleep loss, and patients who underestimate the time they spend asleep may still be suffering from a real sleep deficit as well (Harvey 2002; Harvey & Tang 2012). The distinction between subjective and objective insomnia has also been questioned, as sleep-state misperception may be prevalent in different subtypes of insomnia. As Harvey & Tang (2012) put it, “many patients with insomnia perceive sleep as wake, systematically overestimate the time they take to get to sleep (SOL) and underestimate the time they sleep in total (TST).” This further highlights the urgency of sleep-state misperception existing alongside actual sleep loss in insomnia.

In the context of the present discussion, the example of sleep-state misperception in subjective insomnia may seem to be a counterexample to, rather than a candidate for, dreamless sleep experience. Thompson (2015, p. 5) considers sleep-state misperception as a possible objection to his view: sleep-state misperception

of the type seen in insomnia challenges the reliability of subjective reports of sleep, thus providing a counterexample to his claim that at other times, reports of dreamless sleep experience and of having slept peacefully are veridical memory reports. He then argues that the mere possibility of there being veridical reports of dreamless sleep experience is enough to disprove the default view. He also proposes that in experienced meditators, “we should observe a stronger correlation between subjective reports of phenomenal qualities of sleep and various objective measures of brain activity” (p. 16). The fact that at other times, subjective evaluations of sleep can go wrong does not contradict this view, but merely shows that the investigation of dreamless sleep experience is best restricted to certain subject groups.

Here, I want to suggest that an alternative interpretation of sleep-state misperception is possible. In this alternative view, patients with subjective insomnia are in fact unwitting experts of various kinds of sleep experience. It is merely in conceptualizing their sleep states as occurring in wakefulness that they go wrong. Yet, this is compatible with saying that during sleep, they maintain prereflective awareness of their ongoing sleep state; in fact, it might be their continued perception of sleep that leads them to mischaracterize it as a state of wakefulness, rather than as sleep. Their expertise, consequently, is of a somewhat paradoxical nature: they have a high-degree of familiarity with their sleep, they observe and perhaps even compulsively attend to it—but they don’t recognize or conceptualize it *as* sleep.

Note that this description fits in well with the distinction, introduced at the end of section 5.1, between different readings of the term lucidity. There, I argued that prereflective awareness of the fact that one is now dreaming often precedes the conceptually mediated insight that characterizes full-blown lucidity. Importantly, these two factors may even be dissociable: a fleeting awareness of the dreamlike nature of one’s current state can be misinterpreted, on the level of conscious, conceptually mediated thought, as indicating that one is awake. In such prelucid dreams, the erroneous conclusion

that one is certainly awake may be prompted by the same type of change in experiential character that in other cases drives the cognitive realization that this is a dream (see also Windt 2015, chap. 9).²⁵

Similarly, the idea is that sleep-state misperception arises when patients misinterpret mental activity and phenomenal experience that in fact occurs in sleep as occurring in wakefulness. Indeed, Mercer et al. (2002) found that when they were awakened 5 minutes after the onset of stage 2 sleep or REM sleep, insomnia patients were more likely than good sleepers to say they had been awake. One possibility is that these patients generally have a heightened awareness of sleep-related experiences; another is that increased attention to and concern about the amount of sleep they are getting may increase their sensitivity to such sleep-related experiences, as well as the likelihood of misdescribing them as occurring in wakefulness.

Interestingly, it does not seem that subjective insomnia simply results from a general deficit in the ability to estimate time (Tang & Harvey 2005). Instead, subjective insomnia appears to be associated with selective attention to and increased monitoring of external cues (such as the time of day or the alarm clock), but also of thoughts and bodily sensations that are taken, by the subject, to be inconsistent with sleep. As Mercer et al. (2002, p. 565) put it, “insomniacs’ reduced sleep-wake discriminability may be caused by either a greater amount of mentation during sleep, mentation that more

²⁵ This is also why such examples of prelucid dreams or of sleep-state misperception do not threaten the transparency of retrospective reports. In the present framework, reports are transparent with respect to the occurrence and phenomenal character of experience only; but we should not expect them to accurately reflect the sleep state in which the respective experiences occurred (or indeed whether they occurred in sleep at all), just as we should not expect them to accurately identify the underlying changes in neural activation patterns. Perhaps training, as Thompson (2015) suggests, can indeed improve the match between subjective experience reports and objective measures of sleep states or of brain activation; but this is in no way guaranteed. Or perhaps, objective measures of sleep should be informed by the conditions under which different subject groups experience themselves as being asleep. A mismatch between subjective and objective measures need not indicate a flaw in subjective reports; it might also indicate that objective sleep measures are poorly suited to capture what normal, healthy subjects mean when they say they have been asleep. Indeed, this latter suggestion is in keeping with Thompson’s proposal of a phenomenologically enriched taxonomy of sleep states.

closely resembles awake mentation, or a misattribution of normal nocturnal mentation as wakeful cognitive activity.” Enhanced memory processing may also play a role (Perlis et al. 1997), as might enhanced physiological and cortical arousal. Intriguingly, insomnia patients show heightened beta and gamma EEG activity during sleep onset, but also during NREM sleep; and in one study, this activity was negatively associated with their ability to correctly perceive that they had been asleep (Perlis et al. 2001). Again, this could be an indication of continued awareness during sleep. Subjective insomniacs may be witnessing sleep whilst failing, unlike the expert meditators described by Thompson (2015, see especially p. 16), to realize what it is they are witnessing.

As is the case for white dreams, I am not suggesting that sleep-state misperception in insomnia be equated with dreamless experience, or indeed that any simple explanation is available. Clearly, a wide range of conscious mental activity occurring in sleep might be perceived as occurring in wakefulness, and much of this might be quite different from the specific type of dreamless sleep experience I am interested in here. And equally clearly, sleep-state misperception in insomnia is a far cry from the peaceful type of sleep experience describe in the Indian debate. In her book-length treatment of insomnia, in which she synthesizes research findings with her own personal experience of insomnia, Gayle Greene (2008) describes her reaction to being told, after a sleepless night in the sleep laboratory, that she has in fact been asleep:

So that’s why nobody had come in with a sleeping pill—the EEG said I was asleep. But I was not asleep. I was truly awake. What in the world was it recording? I may have been in a state of deep relaxation, semi-meditative, I usually am when I lie there, and I may have dropped off, but I was aware of all those thoughts, the feel and look of the room, the long drawn-out boredom of lying there without a book to listen to—it felt like consciousness to me. How could I be aware of all that if I hadn’t been awake? (Greene 2008, p. 254)

When asked, however, if she had been aware of the technician coming into her room, she was not (Greene 2008, p. 253).

Here, I want only to make room for the idea that a subgroup of instances of sleep-state misperception might be more properly described as resulting from an awareness of what is in fact sleep, but then is erroneously categorized as belonging to wakefulness. And at least a portion of this awareness of sleep might consist of dreamless sleep experience, or the persistence of temporal experience devoid of further intentional content or any specific objects of awareness during sleep. Moreover, this may well be the dreamless-sleep analogue of prelucid dreaming, where heightened awareness of one’s ongoing state leads to its erroneous characterization as wakeful activity on the level of conceptually mediated thought.

Finally, sleep-state misperception of this type may not be unique to insomnia, but may be prevalent in the general population. In a paper aptly titled “The perceptual uncertainty of having sleep”, Sewitch (1984) describes the outcome of an experiment investigating the ability of healthy subjects to correctly say whether they have been asleep, as determined by objective markers such as EEG measures. She found that out of 210 awakenings from Stage 2 sleep, 116 were judged to be periods of wakefulness; for REM sleep awakenings, the number was lower, with 45 out of 165 awakenings being judged to have been preceded by a period of wakefulness. The surprising conclusion is that even ordinary sleepers quite dramatically underestimate the amount of time they have been asleep (see also Webb 1975). Other studies point in a different direction. There is some evidence that whereas insomnia patients underestimate their total sleep time, healthy subjects overestimate how long they have been asleep (Means et al. 2003; Pinto Jr. et al. 2009). Clearly, more research is needed. But either way, it would seem that our confidence in our ability to tell whether and how long we have been asleep or awake is overrated. And perhaps, at least part of this confusion stems from the fact that the default view is deeply engrained not just in cognitive neuroscience, but also in

folk-psychology: We expect sleep to be a state of unconsciousness, and so when we recall mental activity that is distinct from dreaming, we mistakenly think we must have been awake.

There is, however, an even deeper conceptual point. The mismatch between subjective and objective (behavioral or polysomnographic) markers of sleep should alert us to the fact that conventional definitions of sleep, and attempts to operationalize them scientifically, for instance in the form of sleep-stage scoring systems, may be oversimplified. The borders between sleep and wakefulness themselves may be fluid. This brings us back to Thompson's proposal that a more fine-grained and phenomenologically informed taxonomy of sleep states is needed. This is emphatically illustrated by the following quotation from one of the participants in Sewitch's study. This participant had subjective insomnia and claimed to have been awake following 22 out of 23 Stage 2 NREM sleep awakenings.

Also, there is for me a state which may be technically sleep to you, but is wakefulness to me and, uhh—it's an intermediate state—it's very hard to define, uhh—but I definitely felt that it's there—and uhh—uhh none of the questions precisely examined this situation. (Sewitch 1984, p. 257)

As Thompson suggests, dismantling the default view may be as simple as asking the right kinds of questions.

5.4 Conclusions

I began this commentary by formulating a number of related challenges to Thompson's analysis of dreamless sleep experience. The first two of these centered on the status of reports of dreamless sleep experience. In order to place the scientific investigation of dreamless sleep experience on solid methodological grounding, it is not enough to establish the logical possibility of veridical reports of dreamless sleep experience; rather, some rationale for distinguishing veridical reports from nonveridical ones is needed. Also, in order for such reports to be in-

dicative not just of the occurrence of dreamless sleep experience, but also of its distribution and quantity across sleep, one will have to assume such experiences to be reportable. This means that positive experience reports and reports of an absence of experience, when gathered under the same reporting conditions and unless there is any empirical evidence of disturbing factors, will have to be considered as equally trustworthy. I responded to these dual challenges by pointing out that the methodological background assumptions upon which scientific dream research has long relied, at least implicitly, directly speak to both issues: Dream reports, at least when gathered under (sufficiently) ideal reporting conditions, are indeed assumed to be trustworthy sources of evidence with respect to the occurrence and phenomenal character of experience during sleep (I called this the transparency assumption), and dreams are also assumed to be assumed to be reportable experiences (I called this the reportability assumption). Elsewhere (Windt 2013, 2015), I have argued that both assumptions are theoretically justified because they best explain dream reporting behavior. Here, I only defended the more limited claim that scientific dream research already offers the methodological resources to turn the study of dreamless sleep experience into a scientific research program. This shifts the burden of proof: in order to meaningfully challenge the report-based investigation of dreamless sleep experience, the methodological background assumptions of scientific dream research will have to be challenged as well.

An important upshot was that the default view is inconsistent with scientific dream research. Due to its methodological background assumptions, scientific dream research is committed to the view that if experiences fitting the profile of dreamless sleep experience are, at least occasionally and under sufficiently ideal conditions (for instance immediately after awakening), reported to occur in sleep, then dreamless sleep experience exists. The default view, understood as an *a priori* and conceptually motivated rejection of dreamless sleep experience, is flawed. I then argued that by taking the analogy between contemporary philosoph-

ical and scientific work on dream reports and the Indian debate seriously, valuable lessons can be learned in the other direction as well. In particular, Thompson's reconstruction and critique of the Nyāya syllogism suggests that certain skeptical objections to the trustworthiness of dream reports run into the same problems, resulting either in circularity or an infinite regress.

The second and third parts of my commentary were dedicated to the third challenge to Thompson's view. This was that even if dreamless sleep experience exists, and even if reports of dreamless sleep experience are taken to reflect this fact, its occurrence in experienced meditators is too remote to warrant the large-scale revision of sleep-state taxonomy proposed by Thompson. I attempted to meet this challenge, first, by first sketching the outlines of a conceptual framework for describing dreamless sleep experience. In this framework, dreamless sleep experience is characterized by pure subjective temporality, or the experience of duration and of an extended presence (a stretched phenomenal *now*) arising independently of any further intentional contents, objects of awareness, or modality-specific imagery. This model extends existing work on dreams, where I argue that the simplest forms of dreaming are examples of minimal phenomenal selfhood, or self-location in a spatiotemporal reference frame (Windt 2013, 2015). In dreamless sleep experience, even this minimal form of self-experience is lost; pure subjective temporality during dreamless sleep experience is a candidate for minimal phenomenal experience, or the simplest form of phenomenal consciousness.

In the final part of the commentary, I discussed what I take to be the most plausible candidates for dreamless sleep experience in this sense: these are lucid dreamless sleep, white dreams, and sleep-state misperception as most prominently seen in subjective insomnia. I also proposed that these states can be meaningfully compared to the transition from nonlucid to pre-lucid and fully lucid dreams. Here, my aim was to show that dreamless sleep experience is not a remote possibility, but might plausibly turn out to be a common characteristic of sleep.²⁶

Importantly, I am not claiming that the proposed conceptual model is the final word on dreamless sleep experience; it is only a very first attempt to delineate the borders of the target phenomenon. The model is clearly open to further conceptual refinement, and I would like it understood mainly as an invitation to do so. What I would hope, however, is that the model might facilitate this process by guiding and informing future research. Similarly, the empirical candidates for dreamless sleep experience that I propose should not be taken to be exhaustive, and their plausibility will depend on future research findings. For now, I hope, however, that they lend further support and urgency to Thompson's case for dreamless sleep experience.

Acknowledgments

I would like to thank Roxane Dänner, Martin Dresler, and Valdas Noreika for insightful conversations and helpful comments on an earlier version of this manuscript. And I would like to thank Thomas Metzinger for his guidance, patience, and thoroughly constructive criticism.

less sleep experience, and the addition of dreamless sleep experience to the conceptual tool kit used for the description of sleep and wakefulness, may prove to be no more than a first step in this direction. And while the reconstruction of the Indian debate and its contrast with contemporary views of sleep is a rich and valuable project, important but easily forgotten lessons might be found closer to home as well. The monophasic sleep pattern currently investigated in Western sleep laboratories and taken to be the biological norm may be only a few generations old (cf. Greene 2008, pp. 238-240) and is likely an artifact of a profound change in sleep behavior brought on, to a considerable extent, by electrical lighting. In preindustrial times, sleep was biphasic—two periods of sleep, called the first and the second sleep, were structured around a period of wakefulness that was made up of quiet rest, perhaps even resembling certain meditative states and often involving the contemplation of dreams (Ekirch 2001; Ekirch 2006). Research suggests that under appropriate conditions—in an environment without artificial, electrical lighting and without various nighttime activities that become possible in such an environment, that compete for our attention and increase the pressure for and attraction of staying awake rather than going to bed—we naturally return to this biphasic sleep pattern (Wehr 1992). It does not seem unreasonable to think that the transition to a monophasic sleep pattern, alongside factors such as increased electrical lighting, traffic noise, and time constraints—will have changed not just the structure of sleep, but our experience of sleep as well. With less and less time allotted to sleep, the temptation to simply black out during sleep (or to view sleep as involving such a blackout) may have increased; yet, current sleep behavior in rich, Western societies may be a highly artificial and learned behavior. If we want a taxonomy of sleep states to reflect universal features of sleep, rather than our culturally specific, contemporary sleep habits, we would do well to remember this.

²⁶ Clearly, this is just the very beginning of the conversation on how to refine sleep-state taxonomy. Ultimately, the investigation of dream-

References

- Arstila, V. (Ed.) (2014). *Subjective time: the philosophy, psychology, and neuroscience of temporality*. Cambridge, MA: MIT Press.
- Aserinsky, E. & Kleitman, N. (1953). Regularly occurring periods of eye motility, and concomitant phenomena, during sleep. *Science (New York, N.Y.)*, 118 (3062), 273-274.
- Ayer, A. J. (1960). Professor Malcolm on dreams. *Journal of Philosophy*, 57 (August), 517-535.
- Berkovich-Ohana, A., Dor-Ziderman, Y., Glicksohn, J. & Goldstein, A. (2013). Alterations in the sense of time, space, and body in the mindfulness-trained brain: a neurophenomenologically-guided MEG study. *Consciousness Research*, 4 (912). [10.3389/fpsyg.2013.00912](https://doi.org/10.3389/fpsyg.2013.00912)
- Blanke, O. & Metzinger, T. (2009). Full-body illusions and minimal phenomenal selfhood. *Trends in Cognitive Sciences*, 13 (1), 7-13. [10.1016/j.tics.2008.10.003](https://doi.org/10.1016/j.tics.2008.10.003)
- Bogzaran, F. (2003). Lucid art and hyperspace lucidity. *Dreaming*, 13 (1), 29-42. [10.1023/A:1022186217703](https://doi.org/10.1023/A:1022186217703)
- Brooks, J. E. & Vogelsong, J. A. (2000). *Conscious exploration of dreaming: discovering how we create and control our dreams (Nachdr.)*. Bloomington, Ind: 1st Books Library.
- Cartwright, R. D. (2010). *The twenty-four hour mind: the role of sleep and dreaming in our emotional lives*. Oxford; New York: Oxford University Press.
- Chadha, M. (forthcoming). *Time-Series of Ephemeral Impressions: The Abhidharma-Buddhist View of Conscious Experience, Phenomenology and Cognitive Sciences*. [http://doi.org/10.1007/s11097-014-9354-2](https://doi.org/10.1007/s11097-014-9354-2)
- Conrad, T. (1968). *Zur Wesenslehre des psychischen Lebens und Erlebens: Phenomenologica*. Den Haag: Martinus Nijhoff.
- Dainton, B. (2010). Temporal Consciousness. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy (Spring 2014 Edition)*. <http://plato.stanford.edu/archives/spr2014/entries/consciousness-temporal/>
- Dement, W. & Kleitman, N. (1957). The relation of eye movements during sleep to dream activity: an objective method for the study of dreaming. *Journal of Experimental Psychology*, 53 (5), 339-346.
- Dennett, D. C. (1976). Are dreams experiences? *Philosophical Review*, 73 (April), 151-171.
- Domhoff, G. W. (1996). *Finding Meaning in Dreams*. Boston, MA: Springer US. <http://link.springer.com/10.1007/978-1-4899-0298-6>
- (2003). *The scientific study of dreams: Neural networks, cognitive development, and content analysis*. Washington, DC, US: American Psychological Association. <http://content.apa.org/books/10463-000>
- Dresler, M., Koch, S. P., Wehrle, R., Spoormaker, V. I., Holsboer, F., Steiger, A. & Czigisch, M. (2011). Dreamed movement elicits activation in the sensorimotor cortex. *Current Biology: CB*, 21 (21), 1833-1837. [10.1016/j.cub.2011.09.029](https://doi.org/10.1016/j.cub.2011.09.029)
- Dresler, M., Wehrle, R., Spoormaker, V. I., Koch, S. P., Holsboer, F., Steiger, A. & Czigisch, M. (2012). Neural Correlates of Dream Lucidity Obtained from Contrasting Lucid versus Non-Lucid REM Sleep: A Combined EEG/fMRI Case Study. *Sleep*. [10.5665/sleep.1974](https://doi.org/10.5665/sleep.1974)
- Dresler, M., Erlacher, D., Czigisch, M. & Spoormaker, V. (in press). Lucid dreaming. In M. Kryger, T. Roth & W. Dement (Eds.) *Principles and Practice of Sleep Medicine*.
- Dunlop, C. E. M. (Ed.) (1977). *Philosophical essays on dreaming*. Ithaca: Cornell University Press.
- Ekirch, A. R. (2001). Sleep we have lost: pre-industrial slumber in the British Isles. *The American Historical Review*, 106 (2), 343-386.
- (2006). *At day's close: night in times past*. (1. ed., 1. publ. as a Norton paperback). New York: Norton.
- Fasching, W. (2010). 'I am of the nature of Seeing': Phenomenological Reflections on the Indian Notion of Witness-Consciousness. In M. Siderits, M. E. Thompson & D. Zahavi (Eds.) *Self, No Self?: Perspectives From Analytical, Phenomenological, and Indian Traditions* (pp. 193-216). Oxford: OUP.
- Fink, S. B. (2015). *On Some Foundational Issues in a Neuroscience of Consciousness: Dissociationism, Neural Correlates, and Introspection*. Osnabrück: Dissertation.
- Fink, S. B. (unpublished manuscript). *Introspection and Signal Detection*.
- Flanagan, O. (2001). *Dreaming souls: sleep, dreams, and the evolution of the conscious mind*. (1. Paperback ed). Oxford: Oxford Univ. Press.
- Fox, K. C. R., Zakaraskas, P., Dixon, M., Ellamil, M., Thompson, E. & Christoff, K. (2012). Meditation Experience Predicts Introspective Accuracy. *PLoS ONE*, 7 (9), e45370. [10.1371/journal.pone.0045370](https://doi.org/10.1371/journal.pone.0045370)
- Greene, G. (2008). *Insomniac*. Berkeley: University of California Press.
- Hall, C. S. & Van de Castle, R. L. (1966). *The content analysis of dreams*. New York: Appleton-Century-Crofts.

- Harvey, A. G. (2002). A cognitive model of insomnia. *Behaviour Research and Therapy*, 40 (8), 869-893. [10.1016/S0005-7967\(01\)00061-4](https://doi.org/10.1016/S0005-7967(01)00061-4)
- Harvey, A. G. & Tang, N. K. Y. (2012). (Mis)perception of sleep in insomnia: A puzzle and a resolution. *Psychological Bulletin*, 138 (1), 77-101. [10.1037/a0025730](https://doi.org/10.1037/a0025730)
- Hobson, J. A. (1988). *The Dreaming Brain*. Basic Books.
- Hobson, J. A., Pace-Schott, E. F. & Stickgold, R. (2000). Dreaming and the brain: toward a cognitive neuroscience of conscious states. *The Behavioral and Brain Sciences*, 23 (6), 793-842; discussion 904-1121. [10.1126/science.1149213](https://doi.org/10.1126/science.1149213)
- Horikawa, T., Tamaki, M., Miyawaki, Y. & Kamitani, Y. (2013). Neural Decoding of Visual Imagery During Sleep. *Science*, 340 (6132), 639-642. [10.1126/science.1234330](https://doi.org/10.1126/science.1234330)
- Hoss, R. J. (2010). Content analysis on the potential significance of color in dreams: A preliminary investigation. *International Journal of Dream Research*, 3 (1), 80-90. <http://doi.org/10.11588/ijodr.2010.1.508>
- Hurd, R. (2008). *Exploring the void in lucid dreaming*. <http://dreamstudies.org/2010/05/13/exploring-the-void-in-lucid-dreaming>
- Husserl, E. (2006). *Phantasie und Bildbewusstsein*. Hamburg: F. Meiner.
- Khalsa, S. S., Rudrauf, D., Damasio, A. R., Davidson, R. J., Lutz, A. & Tranel, D. (2008). Interoceptive awareness in experienced meditators. *Psychophysiology*, 45 (4), 671-677. [10.1111/j.1469-8986.2008.00666.x](https://doi.org/10.1111/j.1469-8986.2008.00666.x)
- Kramer, M. (Ed.) (2013). *The Dream Experience: A Systematic Exploration*. New York: Routledge.
- Kühle, L. (2015). Insight-What Is It, Exactly? In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- LaBerge, S. & DeGracia, D. J. (2000). Varieties of lucid dreaming. In R. G. Kunzendorf & B. Wallace (Eds.) *Individual differences in conscious experience* (pp. 269-307). Amsterdam: John Benjamins.
- Leclair-Visonneau, L., Gaymard, B., Leu-Semenescu, S. & Arnulf, I. (2010). Do the eyes scan dream images during rapid eye movement sleep? Evidence from the rapid eye movement sleep behaviour disorder model. *Brain*, 133 (6), 1737-1746. [10.1093/brain/awq110](https://doi.org/10.1093/brain/awq110)
- Le Poidevin, R. (2015). The Experience and Perception of Time. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy (Summer 2015 Edition)*. <http://plato.stanford.edu/archives/sum2015/entries/time-experience/>
- Lutz, A., Lachaux, J.-P., Martinerie, J. & Varela, F. J. (2002). Guiding the study of brain dynamics by using first-person data: Synchrony patterns correlate with ongoing conscious states during a simple visual task. *Proceedings of the National Academy of Sciences*, 99 (3), 1586-1591. [10.1073/pnas.032658199](https://doi.org/10.1073/pnas.032658199)
- Magallón, L. (1991). Awake in the dark: Imageless lucid dreaming. *Lucidity*, 10 (1-2), 46-48.
- Malcolm, N. (1956). Dreaming and Skepticism. *The Philosophical Review*, 65 (1), 14. [10.2307/2182186](https://doi.org/10.2307/2182186)
- (1959a). *Dreaming*. New York: Humanities Press.
- (1959b). Stern's dreaming. *Analysis*, 19 (December)
- McNamara, P., McLaren, D. & Durso, K. (2007). Representation of the Self in REM and NREM Dreams. *Dreaming: Journal of the Association for the Study of Dreams*, 17 (2), 113-126. [10.1037/1053-0797.17.2.113](https://doi.org/10.1037/1053-0797.17.2.113)
- Means, M. K., Edinger, J. D., Glenn, D. M. & Fins, A. I. (2003). Accuracy of sleep perceptions among insomnia sufferers and normal sleepers. *Sleep Medicine*, 4 (4), 285-296. [10.1016/S1389-9457\(03\)00057-1](https://doi.org/10.1016/S1389-9457(03)00057-1)
- Melloni, L. (2015). Consciousness as Inference in Time. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Mercer, J. D., Bootzin, R. R. & Lack, L. C. (2002). Insomniacs' perception of wake instead of sleep. *Sleep*, 25 (5), 564-571.
- Merritt, J. M., Stickgold, R., Pace-Schott, E., Williams, J. & Hobson, J. A. (1994). Emotion Profiles in the Dreams of Men and Women. *Consciousness and Cognition*, 3 (1), 46-60. [10.1006/ccog.1994.1004](https://doi.org/10.1006/ccog.1994.1004)
- Metzinger, T. (2003). *Being no one: the self-model theory of subjectivity*. Cambridge, MA: MIT Press.
- (2013). Why are dreams interesting for philosophers? The example of minimal phenomenal selfhood, plus an agenda for future research1. *Consciousness Research*, 4, 746. [10.3389/fpsyg.2013.00746](https://doi.org/10.3389/fpsyg.2013.00746)
- Metzinger, T. & Windt, J. M. (2014). Die phänomenale Signatur des Wissens: Experimentelle Philosophie des Geistes mit oder ohne Intuitionen? In T. Grundmann, J. Horvath & J. Kipper (Eds.) *Die Experimentelle Philosophie in der Diskussion* (pp. 231-279). Berlin, GER: Suhrkamp.
- (2015). What Does it Mean to Have an Open Mind? In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Murzyn, E. (2008). Do we only dream in colour? A comparison of reported dream colour in younger and older adults with different experiences of black and white media. *Consciousness and Cognition*, 17 (4), 1228-1237. [10.1016/j.concog.2008.09.002](https://doi.org/10.1016/j.concog.2008.09.002)

- Nielsen, T. A. (2000). A review of mentation in REM and NREM sleep: “covert” REM sleep as a possible reconciliation of two opposing models. *The Behavioral and Brain Sciences*, 23 (6), 851-866; discussion 904-1121.
- Noreika, V. (2014). *Alterations in the states and contents of consciousness: Empirical and theoretical aspects. Doctoral thesis*. Turku: Annales Universitatis Turkuensis.
- (2015). It’s not Just About the Contents: Searching for a Neural Correlate of a State of Consciousness. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Noreika, V., Valli, K., Lahtela, H. & Revonsuo, A. (2009). Early-night serial awakenings as a new paradigm for studies on NREM dreaming. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology*, 74 (1), 14-18. [10.1016/j.ijpsycho.2009.06.002](https://doi.org/10.1016/j.ijpsycho.2009.06.002)
- Noreika, V., Windt, J. M., Lenggenhager, B. & Karim, A. A. (2010a). New perspectives for the study of lucid dreaming: From brain stimulation to philosophical theories of self-consciousness. *International Journal of Dream Research*, 3 (1), 36-45. [10.11588/ijodr.2010.1.586](https://doi.org/10.11588/ijodr.2010.1.586)
- Noreika, V., Windt, J. M., Arstila, V., Falter, C. M., Kiverstein, J. D. & Revonsuo, A. (2010b). The subjective and objective duration of static NREM sleep dreams. *International Journal of Dream Research*, 3 (Supplement 1), 6-7.
- Noreika, V., Falter, C. M. & Wagner, T. (2014). Variability of duration perception: From natural and induced alterations to psychiatric disorders. In V. Arstila & D. Lloyd (Eds.) *Subjective Time: The Philosophy, Psychology, and Neuroscience of Temporality* (pp. 529-555). Cambridge, MA: MIT Press.
- Occhionero, M., Cicogna, P., Natale, V., Esposito, M. J. & Bosinelli, M. (2005). Representation of self in SWS and REM dreams. *Sleep and Hypnosis*, 7 (2), 77-83.
- Perlis, M. L., Giles, D. E., Mendelson, W. B., Bootzin, R. R. & Wyatt, J. K. (1997). Psychophysiological insomnia: the behavioural model and a neurocognitive perspective. *Journal of Sleep Research*, 6 (3), 179-188.
- Perlis, M. L., Smith, M. T., Andrews, P. J., Orff, H. & Giles, D. E. (2001). Beta/Gamma EEG activity in patients with primary and secondary insomnia and good sleeper controls. *Sleep*, 24, 110-117.
- Pinto Jr., L. R., Pinto, M. C. R., Goulart, L. I., Truksinas, E., Rossi, M. V., Morin, C. M. & Tufik, S. (2009). Sleep perception in insomniacs, sleep-disordered breathing patients, and healthy volunteers – An important biologic parameter of sleep. *Sleep Medicine*, 10 (8), 865-868. [10.1016/j.sleep.2008.06.016](https://doi.org/10.1016/j.sleep.2008.06.016)
- Pöppel, E. (2003). *Grenzen des Bewußtseins: wie kommen wir zur Zeit, und wie entsteht Wirklichkeit?* (1. Aufl). Frankfurt am Main: Insel-Verl.
- Raffman, D. (1995). On the persistence of phenomenology. In T. Metzinger (Ed.) *Conscious experience*. Paderborn: Schöningh/Imprint Academic.
- Rechtschaffen, A. & Buchignani, C. (1992). *The visual appearance of dreams*.
- Revonsuo, A. (2006). *Inner presence: consciousness as a biological phenomenon*. Cambridge, MA: MIT Press.
- Revonsuo, A., Tuominen, J. & Valli, K. (2015). The Avatars in the Machine. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Rosen, M. G. (2013). What I make up when I wake up: anti-experience views and narrative fabrication of dreams. *Frontiers in Psychology* (4), 514. [10.3389/fpsyg.2013.00514](https://doi.org/10.3389/fpsyg.2013.00514)
- Rosen, M. & Sutton, J. (2013). Self-Representation and Perspectives in Dreams: Perspectives in Dreams. *Philosophy Compass*, 8 (11), 1041-1053. [10.1111/phc3.12082](https://doi.org/10.1111/phc3.12082)
- Schenck, C. (2005). *Paradox lost: midnight in the battleground of sleep and dreams: “violent moving nightmares” (REM sleep behavior disorder), RBD*. [Minneapolis, Minn.]: Extreme-Nights, LLC.
- Schredl, M., Fuchedzhieva, A., Hämig, H. & Schindele, V. (2008). Do we think dreams are in black and white due to memory problems? *Dreaming*, 18 (3), 175-180. [10.1037/1053-0797.18.3.175](https://doi.org/10.1037/1053-0797.18.3.175)
- Schwitzgebel, E. (2002). Why did we think we dreamed in black and white? *Studies in History and Philosophy of Science Part A*, 33 (4), 649-660. [10.1016/S0039-3681\(02\)00033-X](https://doi.org/10.1016/S0039-3681(02)00033-X)
- (2011). *Perplexities of consciousness*. Cambridge, MA: MIT Press.
- Sewitch, D. E. (1984). The Perceptual Uncertainty of Having Slept: The Inability to Discriminate Electroencephalographic Sleep From Wakefulness. *Psychophysiology*, 21 (3), 243-259. [10.1111/j.1469-8986.1984.tb02930.x](https://doi.org/10.1111/j.1469-8986.1984.tb02930.x)
- Siclari, F., Larocque, J. J., Postle, B. R. & Tononi, G. (2013). Assessing sleep consciousness within subjects using a serial awakening paradigm. *Frontiers in Psychology*, 4, 542. [10.3389/fpsyg.2013.00542](https://doi.org/10.3389/fpsyg.2013.00542)
- Siclari, F., LaRocque, J. J., Bernardi, G., Postle, B. R. & Tononi, G. (2014). The neural correlates of consciousness in sleep: a no-task, within-state paradigm. *bioRxiv*, 012443. [10.1101/012443](https://doi.org/10.1101/012443)

- Sikka, P., Valli, K., Virta, T. & Revonsuo, A. (2014). I know how you felt last night, or do I? Self- and external ratings of emotions in REM sleep dreams. *Consciousness and Cognition*, 25, 51-66.
[10.1016/j.concog.2014.01.011](https://doi.org/10.1016/j.concog.2014.01.011)
- Solomonova, E., Fox, K. C. R. & Nielsen, T. (2014). Methodological considerations for the neurophenomenology of dreaming: commentary on Windt's "Reporting dream experience". *Frontiers in Human Neuroscience*, 8, 317. [10.3389/fnhum.2014.00317](https://doi.org/10.3389/fnhum.2014.00317)
- Strauch, I. & Meier, B. (1996). *In search of dreams: results of experimental dream research*. Albany, NY: State University of New York Press.
- Stumbrys, T., Erlacher, D., Schädlich, M. & Schredl, M. (2012). Induction of lucid dreams: A systematic review of evidence. *Consciousness and Cognition*, 21 (3), 1456-1475. [10.1016/j.concog.2012.07.003](https://doi.org/10.1016/j.concog.2012.07.003)
- Sze, J. A., Gyurak, A., Yuan, J. W. & Levenson, R. W. (2010). Coherence between emotional experience and physiology: does body awareness training have an impact? *Emotion (Washington, D.C.)*, 10 (6), 803-814. [10.1037/a0020146](https://doi.org/10.1037/a0020146)
- Tang, N. K. Y. & Harvey, A. G. (2005). Time estimation ability and distorted perception of sleep in insomnia. *Behavioral Sleep Medicine*, 3 (3), 134-150. [10.1207/s15402010bsm0303_2](https://doi.org/10.1207/s15402010bsm0303_2)
- Thompson, E. (2010). *Mind in life: biology, phenomenology, and the sciences of mind*. (1. Harvard Univ. Press paperback ed). Cambridge, Mass.: Belknap Press of Harvard Univ. Press.
- (2014). *Waking, dreaming, being: self and consciousness in neuroscience, meditation, and philosophy*. New York: Columbia University Press.
- (2015). Dreamless Sleep, the Embodied Mind, and Consciousness. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Tononi, G. (2008). Consciousness as Integrated Information: a Provisional Manifesto. *The Biological Bulletin*, 215 (3), 216-242.
- Valli, K., Frauscher, B., Gschliesser, V., Wolf, E., Falkenstetter, T., Schönwald, S. V. & Högl, B. (2012). Can observers link dream content to behaviours in rapid eye movement sleep behaviour disorder? A cross-sectional experimental pilot study. *Journal of Sleep Research*, 21 (1), 21-29. [10.1111/j.1365-2869.2011.00938.x](https://doi.org/10.1111/j.1365-2869.2011.00938.x)
- Voss, U. (2015). Reflections on Insight. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Voss, U., Holzmann, R., Tuin, I. & Hobson, J. A. (2009). Lucid dreaming: a state of consciousness with features of both waking and non-lucid dreaming. *Sleep*, 32 (9), 1191-1200.
- Voss, U., Schermelleh-Engel, K., Windt, J., Frenzel, C. & Hobson, A. (2013). Measuring consciousness in dreams: the lucidity and consciousness in dreams scale. *Consciousness and Cognition*, 22 (1), 8-21. [10.1016/j.concog.2012.11.001](https://doi.org/10.1016/j.concog.2012.11.001)
- Voss, U., Holzmann, R., Hobson, A., Paulus, W., Koppehele-Gossel, J., Klimke, A. & Nitsche, M. A. (2014). Induction of self awareness in dreams through frontal low current stimulation of gamma activity. *Nature Neuroscience*, 17 (6), 810-812. [10.1038/nn.3719](https://doi.org/10.1038/nn.3719)
- Voss, U. & Hobson, A. (2015). What is the State-of-the-Art on Lucid Dreaming? In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Wangyal, T. & Dahlby, M. (1998). *The Tibetan yogas of dream and sleep*. (1st. ed). Ithaca, NY: Snow Lion Publications.
- Webb, W. B. (1975). *Sleep: The gentle tyrant*. Englewood Cliffs, NJ: Prentice Hall.
- Wehr, n. (1992). In short photoperiods, human sleep is biphasic. *Journal of Sleep Research*, 1 (2), 103-107.
- Williford, K. (2015a). Individuation, Integration, and the Phenomenological Subject. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- (2015b). Representationalisms, Subjective Character, and Self-Acquaintance. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Windt, J. M. (2010). The immersive spatiotemporal hallucination model of dreaming. *Phenomenology and the Cognitive Sciences*, 9 (2), 295-316. [10.1007/s11097-010-9163-1](https://doi.org/10.1007/s11097-010-9163-1)
- (2013). Reporting dream experience: Why (not) to be skeptical about dream reports. *Frontiers in Human Neuroscience*, 7, 708. [10.3389/fnhum.2013.00708](https://doi.org/10.3389/fnhum.2013.00708)
- (2015). *Dreaming: a conceptual framework for philosophy of mind and empirical research*. Cambridge, MA; London, UK: MIT Press.
- Windt, J. M., Harkness, D. L. & Lenggenhager, B. (2014). Tickle me, I think I might be dreaming! Sensory attenuation, self-other distinction, and predictive processing in lucid dreams. *Frontiers in Human Neuroscience*, 8. [10.3389/fnhum.2014.00717](https://doi.org/10.3389/fnhum.2014.00717)

- Windt, J. M. & Metzinger, T. (2007). The philosophy of dreaming and self-consciousness: What happens to the experiential subject during the dream state? In D. Barrett & P. McNamara (Eds.) *The new science of dreaming: Vol. 3. Cultural and theoretical perspectives* (pp. 193-248). Westport: Praeger Perspectives.
- Winget, C. N. (1979). *Dimensions of Dreams*. Gainesville: Univ Pr of Florida.
- Wittmann, M. (in press). Modulations of the experience of self and time. *Consciousness & Cognition*
- Wykowska, A. & Arstila, V. (2014). On the flexibility of human temporal resolution. In V. Arstila & D. Lloyd (Eds.) *Subjective Time* (pp. 431-451). Cambridge, MA: MIT Press.

Steps Toward a Neurophenomenology of Conscious Sleep

A Reply to Jennifer M. Windt

Evan Thompson

Windt's groundbreaking commentary expands and enriches my target article by presenting new considerations against the default neuroscience view that "consciousness is that which disappears in dreamless sleep," by proposing a refined conceptual and phenomenological analysis of dreamless sleep experience, and by offering a refined taxonomy of dreamless sleep experiences. These contributions provide new conceptual and methodological tools for the neurophenomenology of sleep and consciousness.

Keywords

Consciousness | Dreamless sleep | Neurophenomenology | Phenomenal selfhood | Self | Time consciousness | Vedānta | Yoga

Author

[Evan Thompson](#)

evan.thompson@ubc.ca

University of British Columbia
Vancouver, BC, Canada

Commentator

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

I would like to begin by thanking Jennifer Windt for her outstanding, constructive commentary ([Windt 2015b](#), this collection) on my target article ([Thompson 2015a](#), this collection), and by expressing my great admiration for her rich discussion, which goes well beyond being a commentary and instead amounts to an original and substantive article in its own right. It is especially gratifying to see the ideas and arguments that I presented be refined and advanced in such a creative and precise way. Indeed, given the wealth of new material that she presents,

her paper calls not so much for a reply as for a commentary of its own. Such a task, however, is beyond the scope of this short reply. Instead, I wish to highlight the advances that Windt makes, so that new experimental research can begin in this area.

The main aims of my target article were (i) to use debates about sleep from classical Indian philosophy to call into question the "default view" in cognitive neuroscience that "consciousness is that which disappears in dreamless sleep," (ii) to suggest instead that there may be

states or phases of dreamless sleep in which consciousness is present, (iii) to argue that sleep science accordingly needs a more refined neurophenomenological taxonomy of sleep states, and (iv) to demonstrate how contemplative methods of mind training provide important resources for the neurophenomenology of sleep and consciousness.

Windt's commentary advances each of these four aims in substantive ways, as I will describe in the following sections.

2 Indian philosophy and sleep science

After answering several possible challenges to my arguments against the default view (see Section 1 of her commentary), Windt shows that the Indian philosophical debate (in which the Yoga and Vedānta schools argue that consciousness persists throughout dreamless sleep, whereas the Nyāya school denies this claim) parallels in certain key respects the Western philosophical and scientific debates about the trustworthiness of dream reports. Given that sleep science must assume as a methodological criterion of dream research that retrospective reports of dreaming and nondreaming are trustworthy (given ideal reporting conditions), we must similarly assume that retrospective reports of the presence or the absence of experience in dreamless sleep are also trustworthy (again, given ideal reporting conditions). This requirement in turn implies that we must refine the conceptual typology of retrospective reports upon awakening from sleep. In Windt's (2015b, p. 11) words, "reports of nondreaming should be further qualified: reporting the absence of experience is not the same as reporting dreamless sleep experience. The former is an instance of reporting an absence of experience, the latter is an instance of reporting a form of experience characterized by the absence of intentional objects; but it is still an experience report." I will not review the steps of her analysis of the methodological requirements of sleep and dream science in detail (see Section 2 of her commentary), but the upshot is that the default view turns out to be inconsistent with the methodological background assumptions of scientific

sleep and dream research. This conclusion strengthens the case against the default view, for whereas I argue that this view is likely to be empirically false, Windt shows that it is inconsistent with the methodological requirements for scientifically investigating the presence and absence of consciousness in sleep.

3 The phenomenology of dreamless sleep experience

Windt's second contribution is to propose a conceptual and phenomenological model of dreamless sleep experience (see Section 3 of her commentary). Starting from my presentation of the Indian conception of dreamless sleep experience as characterized by a feeling of peacefulness and the dissolution of the subject-object duality, as well as my comparison of this conception with Husserl's conception of pre-reflective and pre-egological retentional time consciousness (see Thompson 2007), Windt proposes that dreamless sleep experience is a candidate for minimal phenomenal experience, one characterized only by the phenomenal "now" and a sense of duration, but having no further intentional content. So described, dreamless sleep experience would qualify as the simplest form in which a state can be phenomenally conscious, namely, as minimal phenomenal temporality.

I find this analysis very promising, though two issues require further analysis. The first concerns whether such a minimal phenomenal experience counts as "selfless." Windt proposes that it does, because minimal phenomenal selfhood requires some sense of spatial self-location, whereas dreamless sleep experience consists only in a minimal sense of temporal self-location—not, of course, in the sense of mental time travel (retrospection and prospection), but rather in the sense of a bare feeling of existing "now," with a minimal feeling of flow or duration. Nevertheless, both Advaita Vedānta and Husserl would take issue with this conception of a phenomenal state as "selfless." As I describe in my target article, Advaita Vedānta describes dreamless sleep experience as a state in which the true nature of the self as non-intentional, re-

flexive consciousness is more apparent than in the ordinary waking and dreaming states. For his part, Husserl also describes the pre-egological retentional time consciousness as a minimal structure of self-experience (see [Zahavi 2005](#); [Thompson 2007](#)). It may be that this issue is in part terminological, but there are also likely to be deeper conceptual disagreements about how to analyze the notion of self—whether this notion can be applied to the reflexivity of passive retention (Husserl) or the reflexivity of pure awareness (Vedānta), or whether such states do not meet the criteria for minimal phenomenal selfhood.

Second, and relatedly, I proposed in my target article that, from a Western phenomenological and cognitive scientific perspective, dreamless sleep experience might be describable as a minimal mode of sentience consisting in the feeling of being alive. My point in describing the experience this way was to call attention to the possibly minimal sense of embodiment present in the state. Windt's proposal raises the question of whether even this minimal sense of embodiment may drop away in dreamless sleep, leaving only a bare phenomenal sense of "now." One way to address this question would be to determine whether there can be such a minimal phenomenal temporality in sleep with no affective character, given that one might take the presence of an affective phenomenal character to imply some felt sense of embodiment (assuming that there is a constitutive relation between affect and felt embodiment).

4 The neurophenomenology of sleep states

Windt usefully enlarges the concept of dreamless sleep experience to include a variety of different dreamless sleep states (see Section 4 of her commentary). These states include lucid dreamless sleep (especially the experiential transition from lucid dreaming to lucid dreamless sleep), a possible subclass of white dreams (in which individuals describe the impression of having dreamed but are unable to describe the dream in any detail), subjective insomnia (in which some individuals may maintain pre-re-

flective awareness of their ongoing sleep state while mistakenly conceptualizing their state as wakefulness), in addition to the contemplative practices of lucid dreamless sleep that I describe. Windt's taxonomy is groundbreaking and opens many new avenues for the experimental neurophenomenology of sleep. This is exactly the kind of work I envisioned when I suggested that we need a more fine-grained and phenomenologically informed taxonomy of sleep states.

5 Contemplative sleep states

In my target article, I called attention to the importance of meditative practices of dream yoga and lucid dreamless sleep, because they are closely connected to the Advaita Vedānta, Yoga, and Indian Buddhist conceptions of dreamless sleep, and have begun to be investigated by cognitive neuroscientists (see [Thompson 2015b](#) for further discussion). I agree with Windt that these practices may be too remote from other kinds of sleep experiences in order to justify a wholesale revision of the standard taxonomy of sleep states. For this reason, it is important to place these meditative sleep states within a wider taxonomy that includes other kinds of sleep states, specifically the dreamless sleep states that Windt details. In this way, the meditative practices and their effects on sleep can be integrated into the rest of sleep science. Windt's article provides an excellent framework to this end.

6 Conclusion

Windt's commentary goes far beyond mere commentary in offering new arguments against the default neuroscience view that consciousness is that which disappears in dreamless sleep, by providing a refined conceptual proposal about the phenomenal structure of dreamless sleep experience, and by presenting a new taxonomy of dreamless sleep states and experiences. Thanks to her commentary, sleep science and the neuroscience of consciousness have new conceptual and methodological tools for refining the investigation of consciousness during sleep (see also [Windt 2015a](#)).

References

- Thompson, E. (2007). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Cambridge, MA: Harvard University Press.
- (2015a). Dreamless Sleep, the Embodied Mind, and Consciousness. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- (2015b). *Waking, Dreaming, Being: Self and Consciousness in Neuroscience, Meditation, and Philosophy*. New York: Columbia University Press.
- Windt, J. M. (Ed.) (2015a). *Dreaming: A Conceptual Framework for Philosophy of Mind and Empirical Research*. Cambridge, MA: MIT Press.
- (2015b). Just in Time—Dreamless Sleep Experience as Pure Subjective Temporality – A Commentary on Evan Thompson. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Zahavi, D. (2005). *Subjectivity and Selfhood. Investigating the First-Person Perspective*. Cambridge, MA: MIT Press.

What is the State-of-the-Art on Lucid Dreaming?

Recent Advances and Questions for Future Research

Ursula Voss & Allan Hobson

Lucid dreaming may be defined as the conscious awareness that one is dreaming while dreaming. Instead of incorrectly assuming that one is awake, the dreamer gains insight about her or his real state of consciousness. Lucid dreaming is rare and evanescent, which probably accounts for lingering doubts about its veracity and for its marginalization in science. The purpose of this paper is to review the evidence that lucid dreaming is a real phenomenon, including evidence for its occurrence, underlying mechanisms, and scientific value. Based on admittedly still limited but fast-growing empirical evidence, we will introduce four hypotheses centred around lucid dreaming that are deduced from empirical work and that will hopefully have a bearing on future consciousness research. The Brain Maturation Hypothesis (1) relates steps in ontogenetic brain development to the frequency of naturally occurring lucid dreams in children and adults, suggesting that in the immature brain, spontaneous and involuntary lucid dreaming results from accidental and untypical activation of the frontal cortex during REM sleep. The Hybrid State Hypothesis (2) and the Space of Consciousness Model (SoC) (3) build on the electrophysiological peculiarities observed in REM-sleep-induced lucid dreams, showing a wake-like EEG pattern in frontal parts of the brain and an REM sleep-like EEG in posterior areas. The Gamma Band Hypothesis (4) proposes that the same kind of oscillatory activity known to accompany conscious awareness in the awake brain promotes conscious awareness in REM sleep dreams. Finally, we present first experimental evidence that lower gamma band activity is indeed a necessary condition for the elicitation of conscious awareness in dreams.

Keywords

Brain maturation | Lucid dreaming | REM sleep | Spaces of consciousness model | States of consciousness

Authors

[Ursula Voss](#)

voss@psych.uni-frankfurt.de

Johann Wolfgang Goethe-Universität
Frankfurt a. M., Germany

[Allan Hobson](#)

allan_hobson@hms.harvard.edu

Harvard Medical School
Brookline, Massachusetts, U.S.A.

Commentator

[Lana Kühle](#)

lkuhle@ilstu.edu

Illinois State University
Bloomington-Normal, Illinois, U.S.A.

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Background

Given its robust and revealing features, it is surprising that dream lucidity was not recognized by philosophers of mind until recently ([Metzinger 2003, 1993](#); [Noreika et al. 2010](#); [Revonsuo 2006](#); [Windt in press](#); [Windt & Metzinger 2007](#)). Although it was described by [Aristotle](#) (without using the term, in 350 BC), lucid

dreaming first appears in the *experimental* literature of the late nineteenth century ([Maury 1861](#); [Saint-Denis & Marquis 1982](#)). It was described as a vehicle for self-experimentation ([Arnold-Forster 1921](#)) in the early 20th century and reported on subjectively ([van Eeden 1969](#)). The modern laboratory study of lucid dreaming

1. While dreaming, I was aware of the fact that the things I was experiencing in the dream were not real.	0	1	2	3	4	5
2. While dreaming, I was able to remember my intention to do certain things in the dream.	0	1	2	3	4	5
3. While dreaming, I was aware that the self I experienced in my dream wasn't the same as my waking self.	0	1	2	3	4	5
4. In my dream, I was able to manipulate or control other dream characters in a way that would be impossible in waking.	0	1	2	3	4	5
5. While dreaming, I thought about other dream characters.	0	1	2	3	4	5
6. While dreaming I was able to successfully perform supernatural actions (like flying or passing through walls).	0	1	2	3	4	5
7. The emotions I experienced in my dream were exactly the same as those I would experience in such a situation during wakefulness.	0	1	2	3	4	5
8. While dreaming, I was aware of the fact that the body I experienced in the dream did not correspond to my real sleeping body.	0	1	2	3	4	5
9. I was very certain that the things I was experiencing in my dream wouldn't have any consequences on the real world.	0	1	2	3	4	5
10. While dreaming I was able to successfully control or change the dream environment in a way that would be impossible during wakefulness).	0	1	2	3	4	5
11. While dreaming, I saw myself from outside.	0	1	2	3	4	5
12. While dreaming, I thought about my own actions.	0	1	2	3	4	5
13. While dreaming, I had the feeling that I had forgotten something important.	0	1	2	3	4	5
14. While dreaming, I was able to change or move objects (not persons) in a way that would be impossible in waking.	0	1	2	3	4	5
15. While dreaming I was not myself but a completely different person.	0	1	2	3	4	5
16. While dreaming, I often asked myself whether I was dreaming.	0	1	2	3	4	5
17. The thoughts I had in my dream were exactly the same as I would have in a similar situation during wakefulness.	0	1	2	3	4	5
18. While dreaming, I had the feeling that I could remember my waking life.	0	1	2	3	4	5
19. While dreaming, I was aware of the fact that other dream characters in my dream were not real.	0	1	2	3	4	5
20. Most things that happened in my dream could have also happened during wakefulness.	0	1	2	3	4	5
21. I watched the dream from the outside, as if on a screen.	0	1	2	3	4	5
22. While dreaming, I often thought about the things I was experiencing.	0	1	2	3	4	5
23. I was able to influence the story line of my dreams at will/at libitum.	0	1	2	3	4	5
24. While dreaming, I was able to remember certain plans for the future.	0	1	2	3	4	5
25. While dreaming, I felt euphoric/upbeat.	0	1	2	3	4	5
26. While dreaming, I had strong negative feelings.	0	1	2	3	4	5
27. While dreaming, I had strong positive feelings..	0	1	2	3	4	5
28. While dreaming, I felt very anxious.	0	1	2	3	4	5

Figure 1: Lucidity in Dreams (LuCiD) scale (adopted from [Voss et al. 2013](#)).

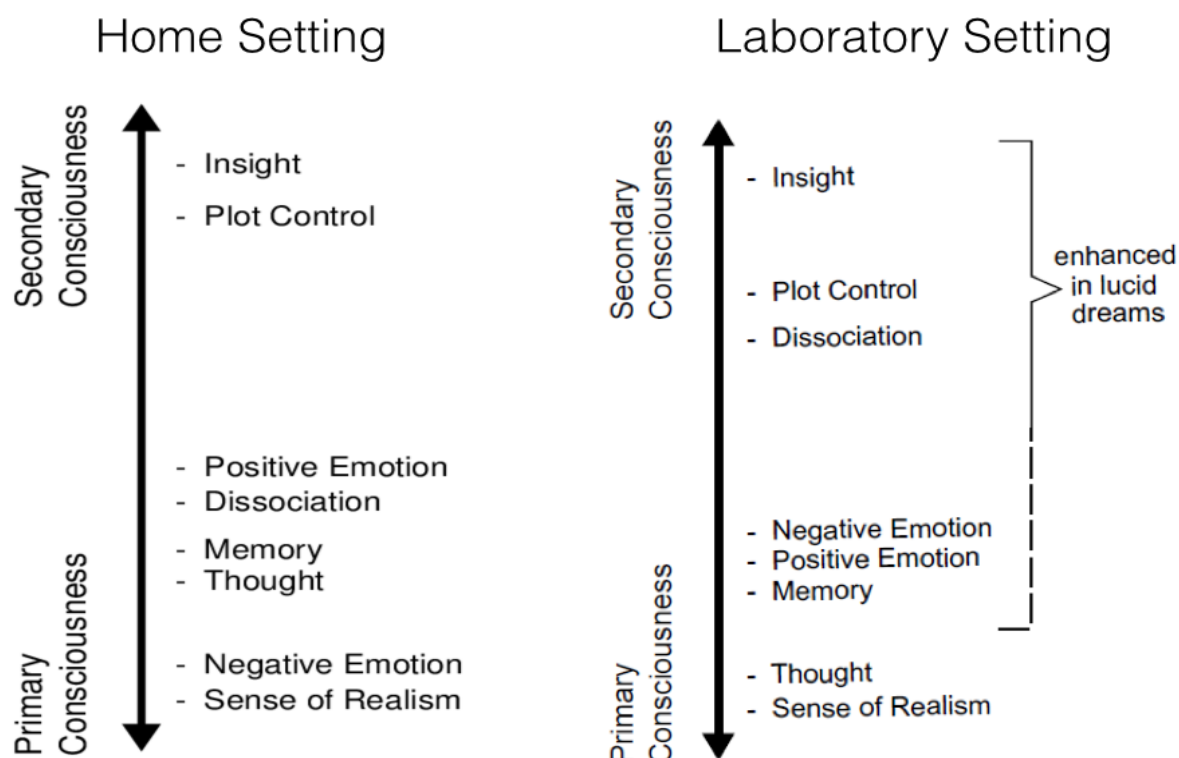


Figure 2: (partially adapted from [Voss et al. 2014](#)): Positions on the primary to secondary consciousness axis are based on the logarithm of ratios of mean scores in lucid and non-lucid dreams. All factors have been identified as components of dream consciousness.

- a) Rank order of logarithm of mean scores derived from dream reports collected in a home setting. Note that these reports were often recorded in the morning instead of immediately following an awakening from REM sleep. Judging from our admittedly limited experience, these reports are less distorted and more story-like than those following forced awakenings in the laboratory.
- b) Rank order of logarithm of mean scores derived from dream reports following forced awakenings from REM sleep in a laboratory setting. Lucid dreams, which are thought to add elements of secondary consciousness, are characterized by increased ratings in reflective INSIGHT, CONTROL over the dream plot, and DISSOCIATION. To a lesser extent, they are accompanied by access to waking MEMORY, as well as NEGATIVE and POSITIVE EMOTIONS.

was pioneered by [Hearne \(1978\)](#) and [LaBerge](#), beginning in 1980 ([1980](#)).

In this paper, we will summarize our five years of scientific research on lucid dreaming, provide a systematic overview of our work, and present new hypotheses about the *why* (because of fluctuations in brain networking) and the *how* (through local changes in lower gamma band activity) of lucid dreaming. Regarding the *why*, our “Brain Maturation Hypothesis” pro-

poses that the probability of lucid dreaming occurring spontaneously is strongly enhanced during the time of cerebral diversification and, most importantly, integration of the frontal lobes into the cortico-cortical and cortico-thalamic networks ([Fuster 1989](#); [Goldman-Rakic 1987](#); [Zilles et al. 1988](#)).

As to the *how* of lucid dreaming, we will outline our experimental findings, focusing on the increase in lower gamma band activity in

fronto-temporal brain areas (Gamma Band Hypothesis). We will then move from our first attempts to provide a brain-based explanation of empirical findings (Hybrid State Hypothesis, see [Hobson & Voss 2011](#); [Voss et al. 2013](#)) to a three-dimensional model of consciousness, allowing for a more structured classification of various states of consciousness, ranging from near-death to highly vigilant wakefulness (“Space of Consciousness”, SoC, compare [Voss & Voss 2014](#)).

The presentation of our empirical contributions will begin with a quantitative analysis of lucid dream subjectivity. This study demonstrates that the most robust difference between lucid and non-lucid dreaming is the increase in insight into the nature of one’s current conscious state that accompanies lucidity. Based on admittedly few recordings (unpublished) of false awakenings, we are currently inclined to assume that there is no notable difference between, for example, the apparent but non-veridical insight accompanying a false awakening and actual (lucid) insight into the fact that one is dreaming. In both falsely and correctly perceived insight into the current state of arousal, the brain apparently operates in a dissociative mode, allowing for a state-related form of meta-awareness similar to the awareness of mind-wandering described for the wake state ([Schooler et al. 2011](#); [Metzinger 2013](#)). However, as our experimentally deduced hypotheses are based on those instances in which the dreamer *correctly* achieved insight into the fact that she was dreaming while the dream continued, we will restrict our discussion of dream lucidity to these instances.

In discussing these results, we will go so far as to suggest that lucidity, as the name implies, *is* insight. We then turn to sleep laboratory studies revealing that the principal brain correlate of lucid dreaming is 40 Hz activation of the frontal cortex. When we electrically stimulated the frontal brain via the scalp, we were able to induce both an increase in 40 Hz brain activation and the subjective experience of lucidity. In our discussion of these results we suggest that the experimental study of lucid dreaming is a powerful paradigm for understanding the brain basis of conscious experience.

2 Quantification of dream lucidity as subjective experience

Perhaps the most problematic aspect of conducting research into lucid dreaming is the difficulty of obtaining both qualified and quantified evidence of the secondary consciousness in dreams. By secondary consciousness we mean the subjective awareness of our state in dreaming, and particularly meta-awareness, meaning an instance of actively acquired self-knowledge or a sudden insight, regardless whether it is accurate or counterfactual (see [Metzinger 2013](#)). Meta-awareness is most clearly manifest in waking consciousness. Dream consciousness, by contrast, is called primary (following [Edelman 1992](#)) because while it is both richly perceptual and powerfully emotional, it is weakly cognitive with conspicuous defects in insight (the main focus of this paper) orientation, and memory, though this does not mean that all thinking is missing ([Hobson et al. 2011](#); [Kahan & Sullivan 2012](#); [Kahn & Hobson 2005](#)). See [Hobson & Voss](#) for detailed discussion of this phenomenology ([2010](#)).

Regarding qualification, [Hearne \(1978\)](#) and [LaBerge \(1980, 1985\)](#) took advantage of the fact that humans can be trained to voluntarily move their eyes in Rapid Eye Movement (REM) sleep and thereby to signal conscious awareness while dreaming. Although care must be taken to minimize the rate of false positive responses, LaBerge’s method has proven quite useful in our own attempts to reliably identify lucid dreaming objectively ([Voss et al. 2009](#)).

With respect to quantification, it is important to note that until recently, lucid dreaming was not quantitatively defined. While some authors described lucid dreams in a narrow sense as dreams in which one knows that one is currently dreaming ([LaBerge 1985](#); [LaBerge & Gackenbach 2000](#)), others subscribed to a broader definition of lucidity as an all-pervading experiential phenomenon, which is purportedly characterized not only by reflective insight into the fact that one is currently dreaming, but also by full intellectual clarity including: the availability of autobiographic memory sources, the ability to actively control the dream, as well as

an overall increase in the intensity of multimodal hallucinatory imagery. This state is often described as taking on a hyper-real quality (Tart 1988; Metzinger 2003; Windt & Metzinger 2007). While sharing an interest in the broader definition, we restrict our attention here to the narrower one in which insight into the fact that one is currently dreaming represents the core criterion for lucidity.

In an attempt to be better able to assess the major and minor determinants of dream lucidity, we developed a Lucidity and Consciousness in Dreams Scale (LuCiD) which was based on hypotheses derived from theory and which we analysed and validated using factor-analysis (Voss et al. 2013). The LuCiD scale presents an important step towards shedding light on the relationship between lucid dreams and other types of dreaming, as well as on the evaluation of cognition in the dream state and its relationship to other aspects of dreaming, such as the intensity of hallucinatory imagery and dream control.

The scale items were constructed by an interdisciplinary team of philosophers, psychiatrists, and psychologists. Our results are based on reports of more than 300 non-lucid and lucid dreams, and verified by reports following forced REM sleep awakenings in the laboratory. Our analysis identified eight factors involved in dream consciousness. Although it is of course possible that our initial item pool did not exhaust all theoretically possible elements, we consider these results a first step in the search for an empirical definition of dream consciousness. According to the factor analysis that we performed, lucid dream consciousness can best be described by the factors (1) INSIGHT into the fact that what one is currently experiencing is not real, but is only a dream; (2) a sense of REALISM, pertaining to the similarity between emotions, thoughts and events with wakefulness as judged after awakening from the dream; (3) CONTROL over the dream plot; (4) access to waking MEMORY; (5) THOUGHT about other dream characters; (6) POSITIVE EMOTION; (7) NEGATIVE EMOTION; and (8) DISSOCIATION akin to taking on a third-person perspective (for a copy of the LuCiD scale see Figure 1).

The factor analysis results support both the restricted definition of lucidity that we have adopted and the broader definition utilised by others. The strength of the factor INSIGHT favors the simple definition, while the wide range of other factors (see Figure 2) favors the more complex definition. While both types of definition certainly have their merits, this difficulty in defining lucid dreams brings some important questions to the fore. What, for instance, is the exact relationship between metacognitive insight into the dream state and the hallucinatory quality of the dream (for the relationship between thinking and hallucinations across the sleep-wake cycle, see Fosse et al. 2001)? And how do these aspects of dream lucidity, in turn, influence the ability to engage in deliberate dream control, which fluctuates considerably?

3 Lucid vs. non-lucid dreams

3.1 Non-lucid dreams

According to our analysis, non-lucid or “normal” dreams are characterized by low absolute values in all factors except REALISM. Non-lucid dreams seem almost to completely lack INSIGHT, CONTROL, and DISSOCIATION. Although mean scores for THOUGHT are higher than those for MEMORY, both are low if we are considering absolute values. Results also show relatively low mean values for NEGATIVE EMOTION. However, as most of our data were collected in a laboratory setting, known to increase positive emotionality in dream imagery (e.g., Hartmann et al. 2001), some caution is advised regarding the interpretation of results with respect to both negative and positive emotion.

3.2 Lucid dreams

Lucid dreams differ from non-lucid dreams in six of the eight factors identified in the LuCiD scale. The leading factor in lucid dreams is INSIGHT. Regarding the relevance of the other factors, we observed different rank orders for dream reports following sleep in a home setting (Figure 2a) and those from forced awakenings in

the laboratory (Figure 2b). The data of our new laboratory study (Voss et al. 2014) confirm the findings depicted in Figure 2b, suggesting that the leading factors in dream lucidity are INSIGHT, CONTROL, and DISSOCIATION. Although, as pointed out by Windt (2013), dream reports in general must be considered trustworthy sources of evidence about subjective experience during sleep, the degree to which these reports can be used to draw scientifically sound conclusions about the dream state strongly depend on the quality of the experimental protocol. Such a protocol is more easily established in a laboratory setting, rendering immediate recalls of the dream experience, which must be considered more reliable with respect to distortions and intermixture with waking thought than those recorded in a home setting (Foulkes 1979; Voss et al., unpublished data), although dreamers might feel less inclined to report on sexual or aggressive content. Furthermore, reports from home settings usually lack information about the particular sleep stage (REM or NREM) in which the dream evolved. Typically, NREM dreams are less bizarre and more story-like (e.g., Dé Waterman & Kenemans 1993).

With regard to the distinction between primary and secondary consciousness in dreams, our findings indicate that INSIGHT is a defining feature of lucidity and that this core aspect of secondary consciousness is related to the emergence of other features of secondary consciousness. Lucid dreamers are able to reflect not only upon the fact that they are currently dreaming, but also upon the unfolding dream events.

The relationship between INSIGHT and CONTROL is clear, as realizing that one is dreaming is an important condition for trying to control not only one's own behavior in the dream, but the dream itself. It must be pointed out, though, that CONTROL is much more infrequent than lucid INSIGHT, and the low covariance of this factor indicates a strongly limited variability of scores, suggestive of a floor effect. In other words, very few participants reported to have experienced some (however small) level of control over the dream plot (see Voss et al. 2013). Despite this limitation, lucid-

ity appears to be characterized not only by lucid insight. INSIGHT also facilitates the emergence of other aspects of secondary consciousness in dreams such as dissociative thought and access to waking MEMORY. Similarly, while our study found non-lucid dreams to almost completely lack INSIGHT, CONTROL, and DISSOCIATION. THOUGHT, e.g., about other dream characters, was not completely absent in non-lucid dreams (Kahn & Hobson 2003).

A surprising finding of our study was that lucid and non-lucid dreams were not distinguished by a difference in the sense of REALISM. Whereas we previously thought that lucidity was characterized by a lack of bizarreness (see Voss et al. 2013), further exploration suggests that this factor is associated with the degree to which the dream feels real. Lucid dreams feel as subjectively realistic as non-lucid dreams. This finding was fully replicated in our most recent study (Voss et al. 2014). A question we are currently not able to answer is whether both dream types are equally bizarre (see also Windt 2013).

Our finding of realistic conviction stands in apparent contrast to reports from other authors who found that the onset of lucidity is often accompanied by a change in the overall experiential quality of the dream, noting that lucid dreams are often described as taking on a surreal, dream-like quality (cf. LaBerge 1985; Brooks & Vogelsson 2000; Tholey & Utecht 2000). At present, we are inclined to think that perhaps the different perceptions may be related to the already-mentioned confounding of wake- and sleep-induced lucid experiences. To our knowledge, lucid dreams entered through the wake state (e.g., Wake-Induced Lucid Dreaming, WILD, see Stumbrys et al. 2012) and those arising out of REM sleep have not been systematically compared with regard to phenomenology or Electroencephalography (EEG). Nonetheless, we think it plausible to assume that the WILD technique will result in more wake-like experiences, simply because they arise out of the wake state or the transition from waking to sleep, usually at the beginning of the night or after morning awakenings. A return to the wake state is in most cases easily ac-

complished. By contrast, dreamers who achieve lucidity out of REM sleep remain in REM sleep, not always being able to wake up voluntarily (Voss et al. 2009, 2014; Voss & Voss 2014). Regarding REALISM, lucid dreams arising out of REM sleep are apparently not accompanied by a change in the subjectively experienced realism of the dream.

3.3 Natural frequency of lucid dreams: The brain maturation hypothesis (1)

REM-sleep-induced lucid dreaming is unique because it represents an exceptional state in which the brain is in two states at the same time: awake and asleep. However, while many have experienced the phenomenon, few experience it on a regular basis. Why? So far, predisposing psychological variables have not been clearly identified (Schredl & Erlacher 2004). We have long speculated (Hobson 2009), and Schredl & Erlacher (2011) have confirmed, that lucid dreaming is negatively correlated with age. Why? And when does lucid dreaming actually set in? These questions need to be addressed in order to provide at least some clues about a very important question: Why does lucid dreaming occur at all?

To investigate the natural frequency of lucid dreaming in children and young adults, we interviewed almost 800 students aged 6–19. Students were recruited from local schools in and around Bonn, Germany, thanks to the enthusiastic cooperation of teachers and parents. Each student was interviewed alone, during school hours, and asked to provide a dream report and to answer questions about dreaming, lucid and non-lucid. In addition, to account for social desirability, students were tested for suggestibility (see Voss et al. 2013), which led to the exclusion of almost 100 data sets.

The main findings of our survey were a surprisingly high incidence of reported lucidity in the young and more frequent lucidity in those who are intellectually more capable. In total, 52% of participating students reported to have recalled at least one lucid episode in their life. The highest incidence rate of recent lucid dreams was observed in the young. Frequency

rates seem to remain at steady levels until age 16, after which they drop dramatically.

In our study, only one third of lucid dreamers claimed to be able to change the dream plot, showing that plot control is not automatically activated in lucid dreaming. As in previous reports (e.g., Wolpin et al. 1992), plot control was significantly associated with frequency of lucid dreaming, suggesting that it is susceptible to training. Plot control was also found to vary with age. It remained at relatively high rates (up to 50% of lucid dreams) from 6 to 14 years and started to decrease from that age on. Lucid dreaming incidence or frequency was not related to sleep duration or napping.

Based on previous research into lucid dreaming, we are inclined to interpret these results as evidence that lucid dreaming is an exceptional mental state occurring naturally in the course of brain maturation. It is noteworthy that the peak in spontaneous occurrence of lucid dreaming coincides with the final stages of frontal lobe myelination and a time of synapse expansion and dendritic growth. These neurobiological changes provide the prerequisites for the integration of the frontal lobes (which are REM sleep-atypically activated in lucid dreaming) into the cortico-cortical and cortico-thalamic networks (Fuster 1989; Goldman-Rakic 1987; Zilles et al. 1988).

Lucid dreaming may thus occur naturally during the final stages of frontal lobe integration, a process similar to an upgrade of computer hardware. It seems to us most likely that the peak in spontaneous dream lucidity in childhood and puberty (Stumbrys et al. 2014; Voss et al. 2013) is nothing but an accidental confounding of conscious states during a time of high cerebral diversification. In an adult, mature brain system, relatively firm covariates for states of arousal have been established. For example, the frontal lobe activity during waking is usually enhanced, whereas it is down-regulated during REM sleep. Our Brain Maturation Hypothesis speculates that during childhood and puberty, frontal lobe activity is sometimes decoupled from the arousal state so that frontal lobes can become active in a state for which this type of activity is untypical. An intriguing

finding is that not only lucid insight but also dissociative phenomena like derealization and depersonalization can easily be trained in the laboratory during this same period in ontogenetic development (Leonard et al. 1999). DISSOCIATION is a key factor that discriminates between lucid and non-lucid dreams (Figure 2, see also van Eeden 1969; Voss et al. 2013, 2014). In lucid dreams, dissociation is often described as taking on a visual third-person perspective, documenting a split between dreamer and dream observer (Gabel 1989; Rossi 1972) (“I see myself from the outside”), whereas non-lucid dreams are typically experienced from the first-person perspective, at least in adults (Foulkes et al. 1990; Gackenbach 2009; Snyder 1970; Voss et al. 2013). At this point, it may be important to note that we do not categorically differentiate between observer dreams and lucid dreams. Based on the results from our LuCiD scale study and in agreement with Gabel (1989), who speaks of “reflections of a dissociated self-monitoring system” (p. 560), we make a quantitative distinction between dreams experienced as first- or third-person, since DISSOCIATION is, next to INSIGHT and plot CONTROL, a key factor that discriminates lucid from non-lucid dreams (see Figure 2).

The fact that lucid dreaming is more readily experienced by those who are more advanced in abstract thinking and charged with reflective insight implies that lucid dreaming is indeed related to brain maturation. Support for this interpretation comes from a study by Lapina et al. (1998). Although details of method and sample characteristics have not been reported, the authors claim a higher level of lucidity in advanced learners. If this is true, however, then why does lucid dreaming decrease in early adulthood, considering that, surely, older students have acquired a higher level of abstraction than younger ones? At this point, we can only speculate about possible and probable causes. One explanation that should be further investigated is that lucid dreaming occurs naturally in the immature but developing brain.

Lucidity could thus be a transient dissociative state during brain maturation that is nor-

mally lost in adulthood but still accessible through training.

3.4 The hybrid state hypothesis (2) of lucid dreaming

The quantification of subjective experience in dream lucidity led us to assume that when the brain-mind shifts from non-lucid to lucid dreaming, it becomes a hybrid state with elements of both waking and dream consciousness. In lucid dreaming, thinking is only partially ruled by primary consciousness. To some extent, the dreamer has—however limited—access to secondary consciousness, enabling her to reflect on her present state. Aside from knowing that the on-going dream is not real, the dream is often experienced as if it were seen from the outside, almost as if the dream were an on-going theatrical production or motion picture (Voss et al. 2014).¹ In other words, lucid dreams can be considered dissociated states of consciousness in which the dream self separates from the on-going flow of mental imagery. The dream is still a dream, but the dreamer is able to distance him or herself from the on-going imagery and may even be successful in gaining (at least partial) control over the dream plot. This phenomenological dissociation is physiologically accompanied by highly selective increases in gamma band activity around 40 Hz in fronto-temporal areas of the brain (Dresler et al. 2012; Voss et al. 2009, 2014), while occipito-parietal regions retain a typical REM-sleep profile. For lucid dreams arising out of REM sleep, we have been able to document the maintenance of sleep throughout the lucid dream, suggesting that lucid dreaming alters REM sleep without surpassing it: REM sleep atonia is unchanged, rapid eye movement bursts (REMs) continue as in REM sleep. However, the EEG frequency spec-

1 We realize that focusing on DISSOCIATION appears to neglect other important aspects of lucid dreaming like agency and knowledge about the ability to exert control, which often seem to occur simultaneously. As a matter of fact, we have observed a significant effect on control, however, during stimulation with 25 Hz but not with 40 Hz, suggesting that oscillatory activity is indeed related to specific brain function. As this is an intriguing but also surprising finding, it is in need of thorough further testing. Please keep in mind that the study of lucid dreaming is still in its fledgling stages and that we have only just begun to explore its possibilities.

trum is significantly altered (Voss et al. 2009). Normally, REM sleep dreams are accompanied by strongly attenuated activation and synchronicity in the gamma frequency band (Castro et al. 2013; Gandal et al. 2012; Voss et al. 2009), especially in frontal parts of the brain (Castro et al. 2013; Voss et al. 2009) suggestive of reduced conscious awareness and executive ego functions (Desmedt & Tomberg 1994). By contrast, gamma band activity in lucid dreaming is significantly increased, while all lower frequencies remain unchanged. This finding strongly suggests that sleep and even REM sleep is indeed maintained. Based on reports of our subjects on their lucid experiences we must assume, however, that lucid dreams push the arousal system towards waking while remaining within the region occupied by REM sleep and thus representing a substate located at the inner boundaries of the REM sleep area within the SoC. Lucid dreaming is, thus, a fragile, destabilized hybrid state. Several participants in our studies have stated that it takes effort to dream lucidly and that such dreams are easily interrupted by noise or state of mind.

Report of a lucid dreamer, f, 30 years old: “To me, being lucid is always a very exciting incident [...] In this state it feels like a struggle in my brain between keeping the dream-scenery and waking. In these short periods of lucidity the awareness of the acting dream body and the real body in bed exist simultaneously and it costs a lot of concentration to keep the balance between both” (for further examples, see Voss & Voss 2014).

We also suggest that lucid dreaming is not just a hybrid state but actually the realization of two normally distinct global functions that usually don't occur simultaneously. This fits in well with the common description of lucid dreams as (partial) awakening in our dreams and involving a split between dreamer and dream-observer, who coexist and change relative dominance of the mind at will (Occhionero et al. 2005). The implications of this line of reasoning have profound impact on the theory of mind. There are two selves, suggesting that the self is a construct elaborated by the brain (Metzinger 2003, 2009, 2013). The two selves of the

lucid dreamer are mediated by distinct brain regions: dreaming is ponto-occipital while lucidity is fronto-cortical. Normally these two brain regions play a winner-takes-all game and dreaming is non-lucid. We come back to this point when we present our physiological model below.

We are attracted by the idea that a key cognitive component of waking, namely insight, can be admixed or even actively injected into REM sleep. Determining the degree to which this enhancement of lucidity is voluntary necessitates a better understanding of altered states of waking conscious awareness, such as hypnosis or mind wandering. We need to know more about conscious state control and to bring that understanding into conjunction with our attempt to understand and influence consciousness.

3.5 Space of Consciousness Model (3)

To speak of lucid dreaming as a hybrid state implies, of course, that states in general have boundaries and intermediates (so-called hybrids). We have, in a recent publication (Voss & Voss 2014) taken this thought further and proposed a model based on the assumption that consciousness is a dynamical process unfolding in a phenomenal state-space continuum occupied by states of arousal such as waking, sleep, and coma. Normally, waking and dreaming constitute two distinct partitions in this state-space. In our new model, what we have called the *hybrid of lucid dreaming* is depicted as a region within the state of REM sleep that stretches REM state variability to the point of destabilizing it, bordering on waking without inducing a complete change of the global configuration.

In our SoC model, we define consciousness as a three-dimensional *space* occupied by *states* that vary as a function of sensing, judging, and motor control. “Sensing” refers to the ability to experience physical and mental fluctuations. “Judging” is meant to describe varying degrees of higher-order cognitive capacities such as reflective awareness, including the ability to dissociate, to think about the past and contemplate the future, and make decisions. The “motor con-

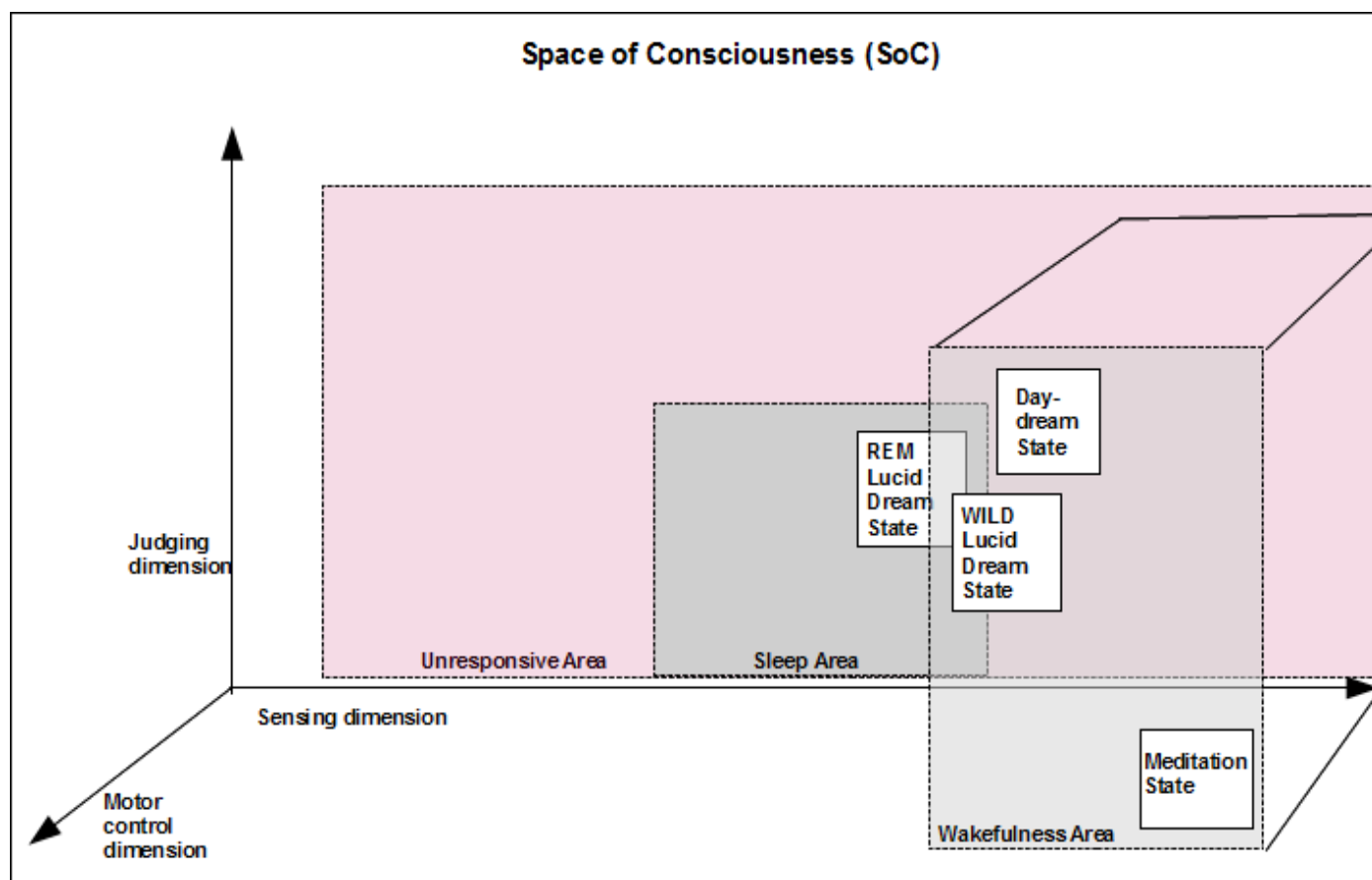


Figure 3: 3-dimensional Space of Consciousness Model (adapted from Voss & Voss 2014, p. 32).

control” dimension was introduced to allow enough space to position different types of unresponsive states such as coma (low motor control, low sensing, and low judging) and, for example, locked-in syndrome, which would be low in motor control but high in sensing and high in judging. Our model is even broad enough to include artificial intelligence (e.g., high judging and low sensing) and to span all forms of animal life as well (see Tononi 2004). Importantly, we do not differentiate between internal and external sources of information or state-dependent neurochemical modulations, as laid out in the AIM model (Hobson et al. 2000; for an early version see Hobson & McCarley 1977). Our space-state model is exclusively phenomenological. The main questions it addresses center around state boundaries and within-state variability.

The *space* is divided into subspaces, corresponding to *states* of arousal, such as waking, sleep, or coma. These *States* largely determine the ability to interact with the external world. We may think of this total space as originating

at the near-death state, spanning over several stages of sleep and wakefulness to some ultimate wake-state of focused attention (see Figure 3). However, it should be kept in mind that the near-death state may not at all be one of minimal expressions of judging and/or sensing (Borjigin et al. 2013; Nelson 2014) so that another altered state may more accurately define the true origin of the SoC.

Lucid dreaming briefly creates a trajectory that dynamically *integrates* the region normally occupied by waking experiences with that of dreaming.

Each state, occupying some area within the SoC, can also be described by a finite number of attributes, and each state possesses a limited degree of variability. Within the partition characterizing wakefulness, for example, we find mind wandering, meditation, and hypnosis, as well as focused attention. Regarding lucid dreams, we assume that wake-induced lucid dreams can be represented by trajectories leading the system very close to the borders, but

which still remain within the overall region inhabited by wake states, whereas REM-sleep-induced lucid dreams initially belong to the sleep state and then evolve towards a brief and unstable integration of the phenomenological substates of waking and dreaming.

Some new questions that derive directly from the model concern (1) the exact number of separable states; (2) specification of the sufficient and causally enabling (perhaps even necessary) conditions allowing for transition from one state into another; and (3) the volume and the dimensionality (the “depth”) of a given region in state-space characterizing each individual state, some perhaps extending over such a broad spectrum of conscious experiences that substates can be defined within the total of SoC and some occupying only a diminutive space such as coma. An example of a high-volume region in phenomenal state-space is wakefulness, covering a wide range of substates including WILD, mind-wandering, focused attention, and hyper-arousal. Another region is sleep, providing a smaller and more dimensionally limited, but nonetheless also considerable range of substates such as light sleep, slow wave sleep, REM sleep (both phasic and tonic), and lucid dreaming.

The SoC model is only an approximation, but we hope that it will prove useful in the generation and testing of specific hypotheses. With regard to lucid dreaming, we hope that this model will contribute to understanding and categorizing the many different aspects and conditions of insightful dreams such as those arising out of the wake state (WILD) versus those arising out of REM sleep. We would expect wake-induced lucid dreams to be accompanied by a much greater motor control, for example, than lucid dreams arising out of REM sleep, simply because WILD are located near the borders of the wake state whereas REM lucid dreams occur in sleep.

3.6 EEG changes associated with lucid dreaming

Our first quantitative EEG study on lucid dreaming aimed to identify changes in brain activity, provided they turned out to be measurable. For this purpose, we trained a relatively

large group of students ($N = 20$) at Bonn University in lucid dreaming. Following several months of preparation, we took those who had achieved lucidity at home 2–3 times per week into the sleep laboratory at the Neurological Clinic of Frankfurt University Hospital.

Although our subjects were highly motivated, our hopes of being able to trace a multitude of lucid dreams soon had to be abandoned, since our enduring attempts yielded EEG recordings of only three spontaneous lucid dreams! Results of this meagre yield were published (Voss et al. 2009), showing that lucid dreaming occurs when activity in the lower gamma band around 40 Hz increases, particularly in frontal parts of the brain. In other words, the results suggested that normal dreaming is cognitively impaired because of frontal lobe deactivation and lucidity only occurs when that deactivation is suspended, either spontaneously or by design.

This finding is depicted in Figure 4, showing single subject 40 Hz EEG power (36–44 Hz) during waking with eyes closed (top), lucid dreaming (middle), and normal non-lucid REM sleep (bottom).

Another finding concerns EEG coherence, or synchronicity (see Figure 5). Whereas the coherence between different brain areas is high in waking (top), it is very low in non-lucid REM sleep (bottom). In lucid dreaming, however, it is significantly increased in comparison to non-lucid dreaming, especially between anterior and posterior parts of the brain (middle).

In this first study, we encountered several methodological difficulties.

1. For the subjects, achieving lucidity in a laboratory setting was difficult. In all three instances, lucid dreaming occurred in the late morning hours, i.e., after 8am. Our study was conducted in the sleep laboratory of the Neurological Clinic at the Frankfurt University Hospital. This implied a noisy early morning routine in which patients were frequently moved for examination purposes, breakfast was served, and floors were cleaned with heavy machinery. It is our opinion now that lucid dreaming arising out of REM sleep

is a fragile state that can be easily disrupted by ambient noise.

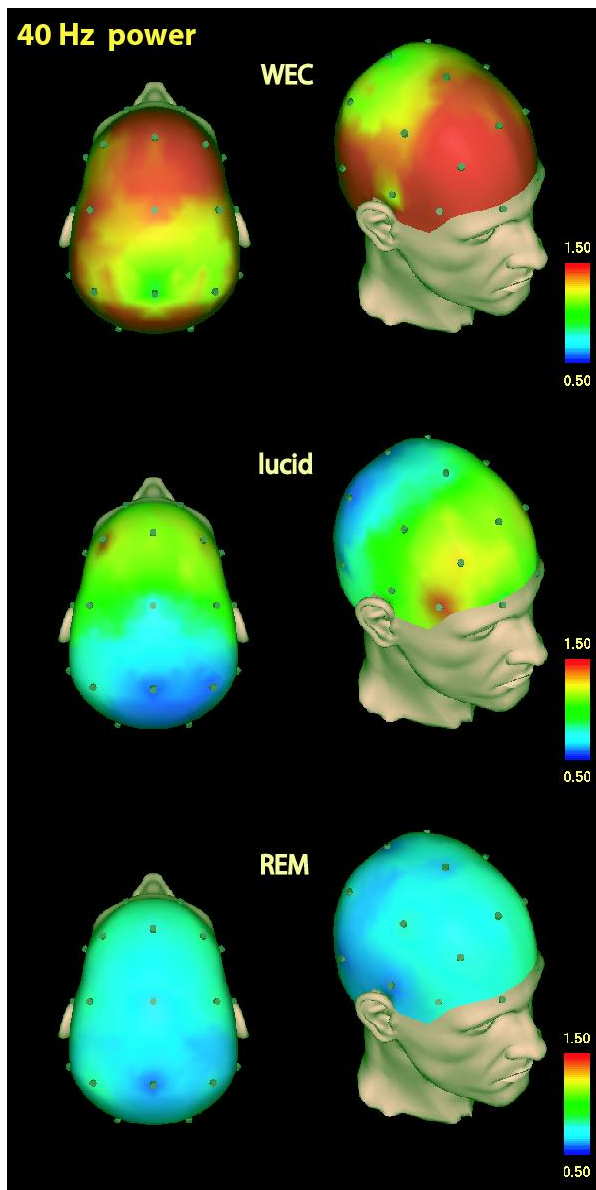


Figure 4: (adapted from Voss et al. 2009). Single subject 40-Hz standardized CSD power during Waking with Eyes Closed (WEC) (top), lucid dreaming (middle), and REM sleep (bottom). Topographic images are based on movement-free EEG episodes and are corrected for ocular artifacts.

2. Several authors have cautioned that some of the variance in gamma activity might be caused by microsaccadic eye movements (Trujillo et al. 2005; Yuval-Greenberg et al. 2008; Weinstein et al. 1991) and by scalp EMG activity (Whitham et al. 2008; Whitham et al. 2007). Although it is not

known, at this point, whether microsaccades are present in steady-states, especially sleep, we have, for publication purposes, conducted a very conservative signal analysis using current source densities (Current Source Densities, CSD). By using this method, we may have overcorrected our EEG scalp potentials, which means that the actual increase in lower gamma band activity is probably even greater than reported.

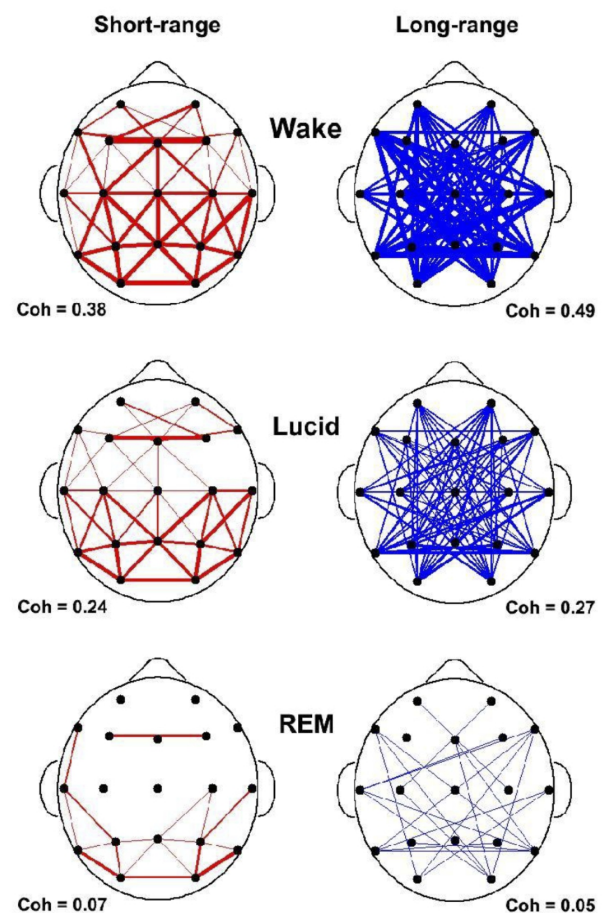


Figure 5: State-dependent short and long range coherences in the 40 Hz frequency band during Waking with Eyes Closed (WEC) (top), lucid dreaming (middle), and non-lucid REM sleep dreaming (bottom). Coherences are indications of interscalp networking and synchronization. Short-range ($N = 55$ pairs) was defined as less than 10cm and long-range (65 pairs) as larger than 15cm inter-electrode distance. Coherences are lowest in REM sleep and strongly enhanced in lucid dreaming.

3. Our subjects reported themselves to be less lucid in the laboratory than at home. When

asked to specify this subjective rating, we found that the subjects' responses were vague and mostly concerned with the amount of plot control achieved in the dream.

The findings of our 2009 study indicate that when subjects became lucid, they shift their EEG power, especially in the 40Hz range and especially in frontal regions of the brain. This shift is, in part, a consequence of pre-sleep auto-suggestion, indicating that REM dream consciousness, which is largely automatic (i.e., spontaneous, involuntary, and intrinsic), is partially subject to volitional force. This observation and its interpretation have an obvious relationship to the question of free will, an implication we will discuss later. Our speculative hypothesis is that dream lucidity arises when wake-like frontal lobe activation is associated with REM-like activity in posterior structures.

3.7 The gamma band hypothesis (4)

In our study of EEG tracings during lucid dreaming, the most striking finding was that lucidity was accompanied by an increased activation of the frontal lobes of the brain. This applies both to synchronicity and to consciousness-related frequencies (around 40 Hz). This observation has led us to propose a “gamma band hypothesis” (Voss et al. 2012; Hobson & Voss 2011), suggesting that brain activation in the 40 Hz frequency range is related to secondary consciousness. We have, in a recent study (Voss et al. 2014), investigated this hypothesis by fronto-temporal application of weak electrical currents in various frequencies. The study was aimed at testing for causality. If activity centered around 40 Hz was causally related to secondary consciousness as expressed in lucid dreaming, then the application of 40 Hz should induce lucid dreaming, provided that it is possible to change brain function in a frequency-specific way through mild electrical stimulation.

3.8 Induction of lucidity via electrical stimulation

In our latest study, we set out to test the hypothesis that lower gamma activity in the

frontal and temporal parts of the brain causally enables lucidity during dreaming. If the observed gamma activity during naturally-occurring lucid dreaming plays a causal role in lucidity, we predicted that facilitation of that frequency band with 40 Hz transcranial alternating current stimulation (tACS) over fronto-temporal areas would increase the probability of lucid dreaming. On the other hand, tACS with a lower or higher frequency should have no effect or even suppress lucid dreaming. The current strength was kept below arousal threshold (250 μ A) in order not to awaken the subjects. Participants were inexperienced lucid dreamers without psychopathology or sleep problems. They were not asked to try to have a lucid dream. Instead, they were told that the study goal was to investigate the effects of mild electrical stimulation in different frequencies on dream content and sleep parameters. While we were doubtful whether it was at all possible to enforce a specific rhythm on the brain (“driving fields”), results suggest that it is indeed possible to change brain activation in a frequency-specific way (see Figure 6). However, we only observed such an effect for frequencies within the lower gamma frequency band. Stimulation with higher or lower frequencies did not result in a measurable change in the respective frequency band, i.e., stimulation with 2 Hz did not lead to an increase in delta frequency band power.

Regarding lower gamma band stimulation, the induced change in lower gamma band brain activity was obviously potent enough to alter conscious awareness in the dream with increased LuCiD ratings especially for INSIGHT and DIS-SOCIATION. Again, this was not observed following stimulation with either higher or lower frequencies.

In this experiment, we tested twenty-seven healthy subjects, during up to four non-consecutive nights. Testing was conducted in a neuro-physiologic sleep laboratory at Goettingen University Hospital. We tested during the summer break of the laboratory and on weekends, which provided a quiet environment and which allowed subjects to continue sleep beyond normal hospital wake-up hours. Participants were allowed

to sleep uninterrupted during the first half of the night until at least 3am.

Starting at 3am, stimulation (30s long) was conducted during REM phases, and subjects were awakened shortly after this stimulation. At this time, they were asked to provide a dream report and ratings to all items of the Lu-CiD scale. The study was performed double blind, so that neither the subject nor the interviewer knew the stimulation frequency applied. In a repeated measures design, all participants were exposed to all stimulation conditions, i.e., sham (no current applied), 2 Hz, 6 Hz, 12 Hz, 25 Hz, 40 Hz, 70 Hz, and 100 Hz (details of methods, see [Voss et al. 2014](#)).

Note that we only applied tACS during REM phases, as lucid dreams arising out of REM sleep were our main research interest. Repetitive stimulation during other sleep stages would have exhausted the experimental protocol and would have led to many undesired side effects such as sleep-deprivation from repetitive early awakenings, changes in sleep architecture, carry-over effects from stimulation in other sleep stages, time-of-night effects, etc.

As shown in Figure 6, only stimulation in the lower gamma band, i.e., stimulation with 25 and 40 Hz, led to an increase in activity in this particular frequency band.

At present, we can only speculate why the other frequencies were not as easily adopted by the brain. Lower frequencies might not have been readily entrained because of state-dependency, as proposed by several authors ([Buzsáki & Draguhn 2004](#); [Vyazovskiy et al. 2009](#); [Tononi et al. 2010](#); [Brown et al. 2012](#); [Suh et al. 2010](#)). It is possible that if we had tried to induce a frequency typically enhanced in slow wave sleep (SWS), for example, such stimulation might have disturbed physiological sleep-dependent oscillations, which would prevent the brain from accepting such a driving field. This notion is supported by direct current (DC) studies (equivalent of 0 Hz) of brain stimulation in REM sleep ([Jakobson et al. 2012a](#); [Stumbrys et al. 2013](#)). Both group of researchers were unable to alter on-going mental activity at 0 Hz, just as we were unable to

induce lucidity at 2, 6, or 12 Hz. Interestingly, dream reports were less frequent in these stimulation conditions ([Voss et al. 2014](#)). However, this does not explain why stimulation with higher frequencies, i.e., 70 and 100 Hz, did not lead to an increase in these frequency bands. It also does not explain why a DC stimulation during stage 2 sleep reportedly effected an increase in visual dream reports although, in this case, the effect was apparently small and, according to the authors, possibly due to arousals and short awakenings ([Jakobson et al. 2012b](#)). At this point, we speculate that lower gamma band frequencies lead to a visible effect because they are linked to the unfolding of secondary consciousness in dreams.

The most striking finding was that subjects reported the ability to “see myself from the outside” and to “watch the dream from the outside as if it was displayed on a screen”. These items belong to the factor DISSOCIATION. Apparently, our subjects took on a third-person perspective following lower gamma band stimulation but not stimulation in any other frequency (2 Hz, 6 Hz, 12 Hz, 70 Hz, 100 Hz) or sham (no current applied).

However, although we were able to induce secondary consciousness in dreams through stimulation with 40 Hz, a similar though smaller effect was observed for stimulation with 25 Hz. Surprisingly, 25 Hz stimulation was associated with CONTROL over the dream plot, whereas stimulation with 40 Hz was not. This finding suggests that specific brain rhythms may be directly linked to cognitive functions and that we have just begun to discover their potential.

Surprisingly, we found no evidence of theta-gamma coupling, as would be expected from NREM sleep studies ([Marshall et al. 2011](#)). At present, we think this may be related to the fact that NREM sleep is highly synchronized, perhaps facilitating such coupling, whereas NREM sleep is desynchronized. As is often the case in science, answering one question generates several others. We will continue to search for answers and also look forward to the extension of our studies by other laboratories.

Effect of tACS on EEG gamma power

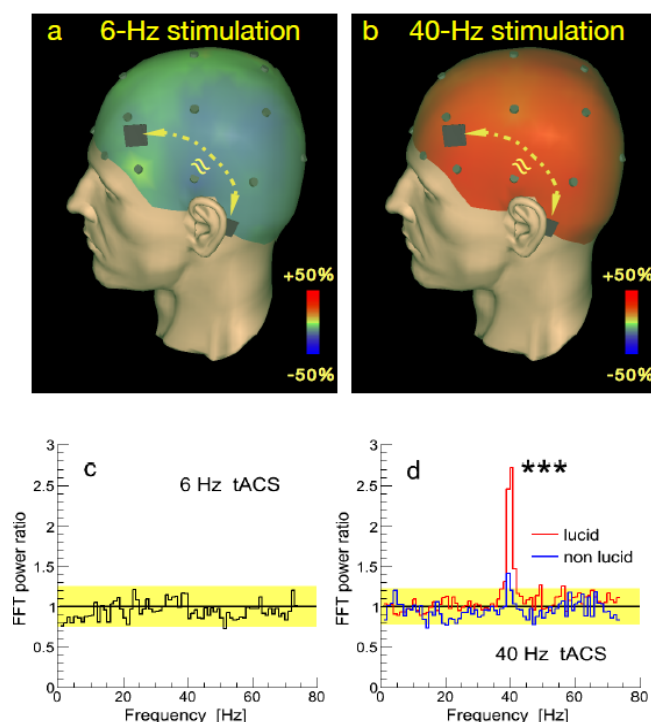


Figure 6: Effect of transcranial alternating current stimulation (tACS) on EEG gamma power. tACS electrodes were placed bilaterally at frontal and temporal positions (black rectangles) and current flowed back and forth between these electrodes. EEG electrode placements are indicated as dark dots.

- Stimulation with 6 Hz resulted in no change in lower gamma activity around 40 Hz (37–43 Hz).
- Stimulation with 40 Hz led to a strong increase in lower gamma activity around 40 Hz.
- Grand average Fast Fourier Transform (Fast Fourier Transform, FFT) power ratios of activity during vs. activity prior to stimulation for the 6 Hz stimulation condition. Yellow shading represents mean values ± 2 standard errors (s.e.). Any excursions outside of this range would be considered significant at least at the $p < .05$ level. However, with 6 Hz, we see no significant stimulation-induced increase in 6 Hz activity.
- Grand average FFT power ratios of activity during vs. activity prior to stimulation for the 40 Hz stimulation condition. Yellow shading represents mean values ± 2 standard errors (s.e.). Note that lucid dreams (red line) are accompanied by a significantly larger increase in the 40 Hz frequency band than non-lucid dreams (blue line) (independent two-sided t tests between lucid and non-lucid dreams during stimulation with 40 Hz: $t_{40\text{Hz}} = 5.01$, $df = 35$, $p < 0.001$).

3.9 Brain Correlates of Lucidity and a Neuropsychological Model.

Our findings of frontal cortical EEG activation to a level intermediate between non-lucid dreaming and waking is compatible with the hybrid state formulation derived from subjective data. More specifically, we attribute the findings to sufficient activation of executive ego functions in the frontal lobes (Baddeley 1992; Goleman & Davidson 1979), but not so intense an activation as to disable the REM sleep generator in the pons and posterior thalamocortical brain that is the physical substrate of dreaming. This formulation is resonant with the oft-repeated complaint that dream lucidity is difficult both to attain and maintain. The hybrid state of waking and dreaming is thus both rare and fragile, suggesting that it is not an adaptive condition for survival and has been eliminated, or reduced to a very low level, by evolution.

It is not difficult to imagine why it would be maladaptive to program waking and dream consciousness at the same time. We will come back to this consideration when we discuss clinical implications below, but at this point we wish to stress the winner-takes-all model that we have sketched as the protoconsciousness hypothesis (Hobson 2009). According to that model, both waking and dreaming are states of consciousness engendered by specifiable brain mechanisms. Waking is governed by aminergic dominance, and dreaming by cholinergic dominance, but both states depend on suppression but not total obliteration of the other. Waking and dreaming are competitive and cooperative brain-mind states.

Of course there is more to the neurophysiology of the differential brain mediation of waking and dreaming. In addition to the chemical neuromodulation mentioned above, we know that REM sleep dreaming is mediated by the active inhibition of both sensory and motor input and output. The data from our studies of lucidity now further suggest that the two states are also differentiated by regional activation of

the cortex. Waking and lucid dreaming are both favored by strong 40 Hz power in the frontal EEG, indicating that frontal lobe activation is a critical mediator of both waking and lucid dream consciousness. Because this sort of activation has been found to correlate with lucidity, we hypothesize that it mediates the wake state component of lucidity. This supposition is also supported by the finding of frontal lobe inactivation in REM sleep, which is correlated with non-lucid dreaming (Braun et al. 1997; Dang-Vu et al. 2007; Desseilles et al. 2011; Nofzinger et al. 1997).

An additional nicety of the theory is that the voluntary eye movements by which lucid dreamers indicate their awareness of their conscious state to third-party observers (Hearne 1978; LaBerge 1980) is evidence of frontal eye field activation in lucid dreamers. This volitional override of the brain stem saccadic eye movement generator is further evidence of the change in the balance of brain-power in several states of consciousness. In lucid dreaming, the wake state control of gaze is returned via frontal lobe activation. According to Metzinger (2013), this is tantamount to the activation of an “epistemic agent model” (EAM), a representation of the self as knowing. This would seem to clinch the argument that conscious states are electrophysiologically differentiated and explained by neurophysiology. This is not surprising, but its specification has been greatly advanced by the scientific investigation of lucid dreaming. A speculative hypothesis that we believe must be tested is that waking entails not only frontal lobe dominance in mediating thought and top-down eye movement control, but that the brain stem itself is primarily harnessed to the analysis of external data with relative suppression of its internal program (see also Activation-Input Gating-Modulation, AIM model, Hobson 1992).

Unfortunately we have no animal model for dream lucidity because we have every reason to suppose that reflective insight such as observed in lucid dreaming necessitates sufficient language capacities assumed essential in the formation of abstract thought (Einstein 1941) or reporting of such. For this reason, we assume that infra-human mammals, which lack significant

language capability, cannot become lucid or report their non-verbal dreams. Whatever one thinks about animal dreams (and we suppose that primary consciousness does accompany their very elaborate REM sleep), no one believes that they are capable of verbally reporting their subjective experience. Dogs and cats do, however, whimper, twitch, and run in their sleep (Lucretius 1995), lending credence to the hypothesis of primary dream-consciousness in animals other than human beings. Animals may dream, and they may become lucid in their dreams, but we doubt the latter and can never offer scientific judgment about either possibility.

The exploration of the physiology of primary consciousness is in its infancy and can be expected to flourish in the future even if we have only rats for subjects (Datta & Hobson 2000; Datta & MacLean 2007). But if we want to learn more about secondary consciousness, we will have to put up with rather severe limitations (Dresler et al. 2012). We trust that advances in brain imaging technology may help this situation. Meanwhile, we hold that the study of lucid dreaming, however difficult, conveys insights about the brain basis of consciousness that is obtainable in no other way.

4 Summary and outlook

What we have learned so far is that the occurrence of lucid dreaming seems to be facilitated by brain maturational processes, in particular the integration of the frontal lobes into the cortico-cortical and cortico-thalamic networks, as outlined in thesis no. 1. Moreover, in lucid dreaming arising out of REM sleep, the apparent spatial dissociation between two states of arousal, waking (rostral) and sleep (caudal) is accompanied by the phenomenological dissociation expressed in an altered conscious awareness, for example, by changing from a first-person to a third-person perspective. This observation has led us to propose that lucid dreaming is to be regarded as a hybrid state (thesis No. 2) within a state-space continuum (thesis No. 3). Another observation concerns changes in frequency-specific oscillatory activity, with significant increases in lower gamma band activity in

lucid dreams, suggesting that lower gamma band activity plays an important role in achieving and/or maintaining a lucid dream. By electrically stimulating the dreaming brain in this frequency band we have been successful in trying to elicit lucid dreams, suggesting a causal role for the gamma frequency band, perhaps not only in lucid dreaming but in higher-order consciousness per se (thesis No. 4).

In spite of this basic scientific progress, our conclusions are only speculative and in need of experimental testing. One future line of research might be the spatial networking involved in consciousness. In our research, we have only stimulated the brain through bilateral fronto-temporal stimulation. We found only lower gamma band activity to be successful in inducing lucid dreaming. What happens, however, when we use different frequencies in rostral and caudal areas? Another question in need of attention is that of applicability. Will wake-training in gamma band activity through Neurofeedback and/or tACS increase the rate of lucid dreaming? What about effects on higher cognitive functions? Finally, we hope that our findings might some day be implemented in clinical settings. This concerns, for example, comatose or locked-in patients who are, through their trauma, confined to a particular state and who may benefit from the possibility of maximally utilizing state capacities.

We have now reviewed and discussed the current state of the art with respect to lucid dreaming. Having been *very* skeptical at first about whether such research could be conducted at all using a rigorous scientific protocol, we have grown increasingly optimistic—if not enthusiastic—about the prospects for the study of lucid dreaming, allowing us to monitor the brain as the mind changes conscious states. In that spirit, lucid dream science may be likened to a moon landing: yes it was hard to achieve, but we did it, and returned to *tell the tale*.

References

- Aristotle, (350 B.C.). *On dreams*.
<http://classics.mit.edu/Aristotle/dreams.html/download>.
- Arnold-Forster, M. (1921). *Studies in dreams*. New York, NY: McMillan.
- Baddeley, A. (1992). Working memory. *Science*, 255 (5044), 556-559. [10.1126/science.1736359](https://doi.org/10.1126/science.1736359)
- Borjigin, J., Lee, T., Liu, T., Pal, D., Huff, S., Klarr, D., Sloboda, J., Hernandez, J., Wang, M. M. & Mashour, G. A. (2013). Surge of neurophysiological coherence and connectivity in the dying brain. *Proceedings of the National Academy of Sciences*, 110 (35), 14432-14437.
- Braun, A. R., Balkin, T. J., Wesenten, N. J., Carson, R. E., Varga, M., Baldwin, P., Selbie, S., Belenky, G. & Herscovitch, P. (1997). Regional cerebral blood flow throughout the sleep-wake cycle. *Brain*, 120 (7), 1173-1197. [10.1093/brain/120.7.1173](https://doi.org/10.1093/brain/120.7.1173)
- Brooks, J. E. & Vogelsong, J. (2000). *The conscious exploration of dreaming: Discovering how we create and control our dreams*. Bloomington, IN: Author House.
- Brown, R., Basheer, R., McKenna, J., Strecker, R. & McCarley, R. (2012). Control of sleep and wakefulness. *Physiological Reviews*, 92 (3), 1087-1187. [10.1152/physrev.00032.2011](https://doi.org/10.1152/physrev.00032.2011)
- Buzsáki, G. & Draguhn, A. (2004). Neuronal oscillations in cortical networks. *Science*, 304 (5679), 1926-1929. [10.1126/science.1099745](https://doi.org/10.1126/science.1099745)
- Castro, S., Falconi, A., Chase, M. & Torterolo, P. (2013). Coherent neocortical 40-Hz oscillations are not present during REM sleep. *European Journal of Neuroscience*, 37 (8), 1330-1339. [10.1111/ejn.12143](https://doi.org/10.1111/ejn.12143)
- Dang-Vu, T. T., Schabus, M., Desseilles, M., Schwartz, S. & Maquet, P. (2007). Neuroimaging of sleep and dreaming. In D. Barrett & P. McNamara (Eds.) *The New Science of Dreaming. Volume 1: Biological Aspects* (pp. 95-114). Westport, CT: Praeger Perspectives.
- Datta, S. & Hobson, J. A. (2000). The rat as an experimental model for sleep neurophysiology. *Behavioral Neuroscience*, 114 (6), 1239-1244. [10.1037/0735-7044.114.6.1239](https://doi.org/10.1037/0735-7044.114.6.1239)
- Datta, S. & MacLean, R. R. (2007). Neurobiological mechanisms for the regulation of mammalian sleep-wake behavior: Reinterpretation of historical evidence and inclusion of contemporary cellular and molecular evidence. *Neuroscience and Biobehavioral Reviews*, 31 (5), 775-824. [10.1016/j.neubiorev.2007.02.004](https://doi.org/10.1016/j.neubiorev.2007.02.004)
- Desmedt, J. & Tomberg, C. (1994). Transient phase-locking of 40 Hz electrical oscillations in prefrontal and parietal human cortex reflects the process of conscious somatic perception. *Neuroscience Letters*, 168 (1-2),

- 126-129. [10.1016/0304-3940\(94\)90432-4](https://doi.org/10.1016/0304-3940(94)90432-4)
- Desseilles, M., Dang-Vu, T. T., Sterpenich, V. & Schwartz, S. (2011). Cognitive and emotional processes during dreaming: a neuroimaging view. *Consciousness and Cognition*, 20 (4), 998-1008. [10.1016/j.concog.2010.10.005](https://doi.org/10.1016/j.concog.2010.10.005)
- Dresler, M., Wehrle, R., Spoormaker, V. I., Koch, S. P., Holsboer, F., Steiger, A., Obrig, H., Sämann, P. G. & Czisch, M. (2012). Neural correlates of dream lucidity obtained from contrasting lucid versus non-lucid REM sleep: a combined EEG/fMRI case study. *Sleep*, 35 (7), 1017-1020. [10.5665/sleep.1974](https://doi.org/10.5665/sleep.1974)
- Dé Waterman, M. E., Elton, M. & Kenemans, J. L. (1993). Methodological issues affecting the collection of dreams. *Journal of Sleep Research*, 2 (1), 8-12. [10.1111/j.1365-2869.1993.tb00053.x](https://doi.org/10.1111/j.1365-2869.1993.tb00053.x)
- Edelman, G. M. (1992). *Bright air, brilliant fire: On the matter of the mind*. New York, NY: Basic Books.
- Einstein, A. (1941). The common language of science. *Out of my later years*. Radio Recording: British Association for the Advancement of Science.
- Fosse, R., Stickgold, R. & Hobson, J. A. (2001). Brain-mind states: Reciprocal variations in thoughts and hallucinations. *Psychological Science*, 12 (1), 30-36. [10.1111/1467-9280.00306](https://doi.org/10.1111/1467-9280.00306)
- Foulkes, D. (1979). Home and laboratory dreams - 4 empirical studies and a conceptual re-evaluation. *Sleep*, 2 (2), 233-251.
- Foulkes, D., Hollifield, M., Sullivan, B., Bradley, L. & Terry, R. (1990). REM dreaming and cognitive skills at ages 5-8: A cross-sectional study. *International Journal of Behavioral Development*, 13 (4), 447-465.
- Fuster, J. M. (1989). *The prefrontal cortex*. New York, NY: Raven.
- Gabel, S. (1989). Dreams as a possible reflection of a dissociated self-monitoring system. *The Journal of nervous and mental disease*, 177 (9), 560-568. [10.1097/00005053-198909000-00008](https://doi.org/10.1097/00005053-198909000-00008)
- Gackenbach, J. I. (2009). Video game play and consciousness development: A replication and extension. *International Journal of Dream Research*, 2 (1), 3-11. [10.1037/1053-0797.16.2.96](https://doi.org/10.1037/1053-0797.16.2.96)
- Gandal, M., Edgar, J. & Klook K., Siegel S. (2012). Gamma synchrony: Towards a translational biomarker for the treatment-resistant symptoms of schizophrenia. *Neuropharmacology*, 62 (3), 1504-1518. [10.1016/j.neuropharm.2011.02.007](https://doi.org/10.1016/j.neuropharm.2011.02.007)
- Goldman-Rakic, P. (1987). Cerebral cortical mechanisms in schizophrenia. *Neuropsychopharmacology*, 10 (3), 22-27.
- Goleman, D. & Davidson, R. (1979). *Consciousness: Brain, states of awareness, and mysticism*. New York, NY: Harper and Row.
- Hartmann, E., Zborowski, M. & Kunzendorf, R. (2001). The emotion pictured by a dream: An examination of emotions contextualized in dreams. *Sleep and Hypnosis*, 3 (1), 33-43.
- Hearne, K. (1978). *Lucid dreams: An electro-physiological and psychological study*. Liverpool, UK: University of Liverpool, England.
- Hobson, J. A. (1992). A new model of the brain-mind state: Activation level, input source, and mode of processing (AIM). In J. S. Antrobus & M. Bertini (Eds.) *Neuropsychology of sleep and dreaming* (pp. 227-245). Hillsdale, NJ: Lawrence Erlbaum.
- (2009). REM sleep and dreaming: towards a theory of protoconsciousness. *Nature Reviews Neuroscience*, 10 (11), 803-813. [10.1038/nrn2716](https://doi.org/10.1038/nrn2716)
- Hobson, J. A., Pace-Schott, E. F. & Stickgold, R. (2000). Dreaming and the brain: Toward a cognitive neuroscience of conscious states. *Behavioral and Brain Sciences*, 23 (6), 793-842.
- Hobson, J. A., Sangsanguan, S., Arantes, H. & Kahn, D. (2011). Dream logic – The inferential reasoning paradigm. *Dreaming*, 21 (1), 1-15. [10.1037/a0022860](https://doi.org/10.1037/a0022860)
- Hobson, J. A. & McCarley, R. W. (1977). The brain as a dream state generator: An activation- synthesis hypothesis of the dream process. *American Journal of Psychology*, 134 (12), 1335-1348.
- Hobson, J. A. & Voss, U. (2010). Lucid dreaming and the bimodality of consciousness. *Towards new horizons in consciousness research from the boundaries of the brain* (pp. 155-165). John Benjamins: Amsterdam, NL.
- (2011). A mind to go out of: Reflections on primary and secondary consciousness. *Consciousness and Cognition*, 20 (4), 993-997. [10.1016/j.concog.2010.09.018](https://doi.org/10.1016/j.concog.2010.09.018)
- Jakobson, A., Conduit, R. & Fitzgerald, P. B. (2012a). Investigation of visual dream reports after transcranial direct current stimulation (tDCS) during REM sleep. *International Journal of Dream Research*, 5 (1), 87-93. [10.11588/ijodr.2012.1.9272](https://doi.org/10.11588/ijodr.2012.1.9272)
- Jakobson, A. J., Fitzgerald, P. B. & Conduit, R. (2012b). Induction of visual dream reports after transcranial direct current stimulation (tDCs) during stage 2 sleep. *Journal of Sleep Research*, 21 (4), 369-379. [10.1111/j.1365-2869.2011.00994.x](https://doi.org/10.1111/j.1365-2869.2011.00994.x)
- Kahan, T. L. & Sullivan, K. T. (2012). Assessing metacognitive skills in waking and sleep: A psychometric analysis of the metacognitive, affective, cognitive experience (MACE) questionnaire. *Consciousness and Cognition*, 21

- (1), 340-352. [10.1016/j.concog.2011.11.005](https://doi.org/10.1016/j.concog.2011.11.005).
- Kahn, D. & Hobson, J. A. (2003). State dependence of character perception - Implausibility differences in dreaming and waking consciousness. *Journal of Consciousness Studies*, 10 (3), 57-68.
- (2005). State-dependent thinking: A comparison of waking and dreaming thought. *Consciousness and Cognition*, 14 (3), 429-438. [10.1016/j.concog.2004.10.005](https://doi.org/10.1016/j.concog.2004.10.005).
- LaBerge, S. P. (1980). Lucid dreaming as a learnable skill – a Case Study. *Perceptual and Motor Skills*, 51 (3), 1039-1042. [10.2466/pms.1980.51.3f.1039](https://doi.org/10.2466/pms.1980.51.3f.1039)
- (1985). *Lucid Dreaming*. New York, NY: Ballantine Books.
- LaBerge, S. & Gackenbach, J. (2000). Lucid dreaming. In E. Cardena, S. J. Lynn & S. Krippner (Eds.) *Varieties of anomalous experience: Examining the scientific evidence* (pp. 151-183). Washington, DC: American Psychological Association.
- Lapina, N., Lysenko, V. & Burikov, A. (1998). Age-dependent dreaming characteristics of secondary schools pupils. *Sleep*, 1, 287-287.
- Leonard, K., Telch, M. & Harrington, P. (1999). Dissociation in the laboratory: A comparison of strategies. *Behavior Research and Therapy*, 37 (1), 49-61. [10.1016/S0005-7967\(98\)00072-2](https://doi.org/10.1016/S0005-7967(98)00072-2)
- Lucretius, (1995). *On the nature of things (De rerum natura)*. Baltimore, MD: Johns Hopkins Press.
- Marshall, L., Kirov, R., Brade, J., Mölle, M. & Born, J. (2011). Transcranial electrical currents to probe EEG brain rhythms and memory consolidation during sleep in humans. *PLoS One*, 6 (2), e16905-e16905. [10.1371/journal.pone.0016905](https://doi.org/10.1371/journal.pone.0016905)
- Maury, A. (1861). *Le sommeil et les rêves*. Paris, FR: Didier.
- Metzinger, T. (1993). *Subjekt und Selbstmodell. Die Perspektivität phänomenalen Bewusstseins vor dem Hintergrund einer naturalistischen Theorie mentaler Repräsentation*. Paderborn, GER: mentis.
- (2003). *Being no one: The self-model theory of subjectivity*. Cambridge, MA: MIT Press.
- (2009). *The ego tunnel: The science of the mind and the myth of the self*. New York, NY: Basic Books.
- (2013). Why are dreams interesting for philosophers? The example of minimal phenomenal selfhood, plus an agenda for future research. *Frontiers in Psychology*, 4, 1-17. [10.3389/fpsyg.2013.00746](https://doi.org/10.3389/fpsyg.2013.00746)
- (2013). The myth of cognitive agency: Subpersonal thinking as a cyclically recurring loss of mental autonomy. *Frontiers in Psychology*, 4 (931). [10.3389/fpsyg.2013.00931](https://doi.org/10.3389/fpsyg.2013.00931)
- Nelson, K. R. (2014). Near-death experience: Arising from the borderlands of consciousness in crisis. *Annals of the New York Academy of Sciences*, 1330 (1), 111-119.
- Nofzinger, E. A., Mintun, M. A., Wiseman, M. B., Kupfer, D. J. & Moore, R. Y. (1997). Forebrain activation in REM sleep: An FDG PET study. *Brain Research*, 770 (1-2), 192-201. [10.1016/S0006-8993\(97\)00807-X](https://doi.org/10.1016/S0006-8993(97)00807-X)
- Noreika, V., Windt, J. M., Lenggenhager, B. & Karim, A. A. (2010). New perspectives for the study of lucid dreaming: From brain stimulation to philosophical theories of self-consciousness. Commentary on “The neurobiology of consciousness: Lucid dreaming wakes up” by J. Allan Hobson. *International Journal of Dream Research*, 3 (1), 36-46.
- Occhionero, M., Cicogna, P., Natale, V., Esposito, M. J. & Bosinelli, M. (2005). Representation of self in SWS and REMdreams. *Sleep and Hypnosis*, 7 (2), 77-83.
- Revonsuo, A. (2006). *Inner presence: Consciousness as a biological phenomenon*. Cambridge, MA: MIT Press.
- Rossi, E. L. (1972). Self-reflection in dreams. *Psychotherapy: Theory, Research & Practice*, 9 (4), 290-298. [10.1037/h0086773](https://doi.org/10.1037/h0086773)
- Saint-Denis, D. H. de & Marquis, J. M. L. (1982). *Dreams and the means of directing them*. London, UK: Gerald Duckworth.
- Schooler, W., Smallwood, J., Christoff, K., Handy, T. C., Reichle, E. D. & Sayette, M. A. (2011). Meta-awareness, perceptual decoupling and the wandering mind. *Trends in Cognitive Sciences*, 15 (7), 317-326. [10.1016/j.tics.2011.05.006](https://doi.org/10.1016/j.tics.2011.05.006)
- Schredl, M. & Erlacher, D. (2004). Lucid dreaming frequency and personality. *Personality and Individual Differences*, 37 (7), 1463-1473. [10.1016/j.paid.2004.02.003](https://doi.org/10.1016/j.paid.2004.02.003)
- (2011). Lucid dreaming frequency in a representative German sample. *Percept Motor Skill*, 112 (1), 104-108. [10.2466/09.PMS.112.1.104-108](https://doi.org/10.2466/09.PMS.112.1.104-108)
- Snyder, F. (1970). The phenomenology of dreaming. In L. Madow & L. H. Snow (Eds.) *The psychodynamic implications of the physiological studies on dreams* (pp. 124-151). Springfield, IL: C. C. Thomas.
- Stumbrys, T., Erlacher, D., Schädlich, M. & Schredl, M. (2012). Induction of lucid dreams: a systematic review of evidence. *Consciousness and Cognition*, 21 (3), 1456-1475. [10.1016/j.concog.2012.07.003](https://doi.org/10.1016/j.concog.2012.07.003)
- Stumbrys, T., Erlacher, D. & Schredl, M. (2013). Testing the involvement of the prefrontal cortex in lucid dreaming: A tDCS study. *Consciousness and Cognition*, 22 (4), 1214-1222. [10.1016/j.concog.2013.08.005](https://doi.org/10.1016/j.concog.2013.08.005)
- Stumbrys, T., Erlacher, D., Johnson, M. & Schredl, M. (2014). The phenomenology of lucid dreaming: An online survey.

- American Journal of Psychology*, 127 (2), 191-204.
- Suh, H. S., Lee, W. H., Cho, Y. S., Kim, J. H. & Kim, T. S. (2010). Reduced spatial focality of electrical field in tDCS with ring electrodes due to tissue anisotropy. *Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society.*, 2053-2056. [10.1109/IEMBS.2010.5626502](https://doi.org/10.1109/IEMBS.2010.5626502)
- Tart, C. (1988). From spontaneous event to lucidity - A review of attempts to consciously control nocturnal dreaming. In J. Gackenbach & S. LaBerge (Eds.) *Conscious mind, sleeping brain* (pp. 67-103). New York, NY: Plenum Press.
- Tholey, P. & Utecht, K. (2000). *Schöpferisch träumen: Wie Sie im Schlaf das Leben meistern. [creative dreaming: how you can master your life while dreaming.]*. Eschborn, GER: Klotz.
- Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, 5, 42.
- Tononi, G., Riedner, B., Hulse, B., Ferrarelli, F. & Sarasso, S. (2010). Enhancing sleep slow waves with natural stimuli. *Medicamundi*, 54 (2), 73-79.
- Trujillo, L. T., Peterson, M. A., Kaszniak, A. W. & Allen, J. J. B. (2005). EEG phase synchrony differences across visual perception conditions may depend on recording and analysis methods. *Clinical Neurophysiology*, 116 (1), 172-189. [10.1016/j.clinph.2004.07.025](https://doi.org/10.1016/j.clinph.2004.07.025)
- van Eeden, F. (1969). A study of dreams. In C. Tart (Ed.) *Altered states of consciousness* (pp. 145-157). New York, NY: Wiley.
- Voss, U., Holzmann, R., Tuin, I. & Hobson, J. A. (2009). Lucid dreaming: A state of consciousness with features of both waking and non-lucid dreaming. *Sleep*, 32 (9), 1191-1200.
- Voss, U., Frenzel, C., Koppehele-Gossel, J. & Hobson, J. A. (2012). Lucid dreaming: An age dependent brain dissociation. *Journal of Sleep Research*, 21 (6), 634-642. [10.1111/j.1365-2869.2012.01022.x](https://doi.org/10.1111/j.1365-2869.2012.01022.x)
- Voss, U., Schermelleh-Engel, K., Windt, J., Frenzel, C. & Hobson, J. A. (2013). Measuring consciousness in dreams: The lucidity and consciousness in dreams scale. *Consciousness and Cognition*, 22 (1), 8-21. [10.1016/j.concog.2012.11.001](https://doi.org/10.1016/j.concog.2012.11.001)
- Voss, U., Holzmann, R., Hobson, A., Paulus, W., Koppehele-Gossel, J., Klimke, A. & Nitsche, M. (2014). Induction of self awareness in dreams through frontal low current stimulation of gamma activity. *Nature Neuroscience*, 17 (6), 810-812. [10.1038/nn.3719](https://doi.org/10.1038/nn.3719)
- Voss, U. & Voss, G. (2014). A neurobiological model of lucid dreaming. *Lucid dreaming: New perspectives on consciousness in sleep* (pp. 23-26). Santa Barbara, CA: Praeger.
- Vyazovskiy, V., Faraguna, U., Cirelli, C. & Tononi, G. (2009). Triggering slow waves during NREM sleep in the rat by introcortical electrical stimulation: Effects of sleep/wake history and background activity. *Journal of Neurophysiology*, 101 (4), 1921-1931. [10.1152/jn.91157.2008](https://doi.org/10.1152/jn.91157.2008)
- Weinstein, J. M., Balaban, C. D. & Verl-Hoeve, J. N. (1991). Directional tuning of the human presaccadic spike potential. *Brain Research*, 543 (2), 243-250. [10.1016/0006-8993\(91\)90034-S](https://doi.org/10.1016/0006-8993(91)90034-S)
- Whitham, E. M., Pope, K. J., Fitzgibbon, S. P., Lewis, T., Clark, C. R., Loveless, S., Broberg, M., Wallace, A., DeLosAngeles, D., Lillie, P., Hardy, A., Fronsco, R., Pulbrook, A. & Willoughby, J. O. (2007). Scalp electrical recording during paralysis: Quantitative evidence that EEG frequencies above 20 Hz are contaminated by EMG. *Clinical Neurophysiology*, 118 (8), 1877-1888. [10.1016/j.clinph.2007.04.027](https://doi.org/10.1016/j.clinph.2007.04.027)
- Whitham, E. M., Lewis, T., Pope, K. J., Fitzgibbon, S. P., Clark, C. R., Loveless, S., DeLosAngeles, D., Wallace, A. K., Broberg, M. & Willoughby, J. O. (2008). Clinical Neurophysiology. *Thinking activates EMG in scalp electrical recordings.*, 119 (5), 1166-1175. [10.1016/j.clinph.2008.01.024](https://doi.org/10.1016/j.clinph.2008.01.024)
- Windt, J. (2013). Reporting dream experience: Why (not) to be skeptical about dream reports. *Frontiers In Human Neuroscience*, 7 (708). [10.3389/fnhum.2013.00708](https://doi.org/10.3389/fnhum.2013.00708).
- (in press). *Dreaming. A conceptual framework for philosophy of mind and empirical research*. Boston, MA: MIT Press.
- Windt, J. & Metzinger, T. (2007). The philosophy of dreaming and self-consciousness: What happens to the experiential subject during the dream state? In D. Barrett & P. McNamara (Eds.) *The new science of dreaming, Vol. 3: Cultural and theoretical perspectives* (pp. 193-247). Westport, CT: Praeger Perspectives/Greenwood Press.
- Wolpin, M., Marston, A., Randolph, C. & Clothier, A. (1992). Individual difference correlates of reported lucid dreaming frequency and control. *Journal of Mental Imagery*, 16 (3-4), 231-236.
- Yuval-Greenberg, S., Tomer, O., Keren, A. S., Nelken, I. & Deouell, L. Y. (2008). Transient induced gamma-band response in EEG as a manifestation of miniature saccades. *Neuron*, 58 (3), 429-441. [10.1016/j.neuron.2008.03.027](https://doi.org/10.1016/j.neuron.2008.03.027)
- Zilles, K., Armstrong, E., Schleicher, A. & Kretschmann, H. J. (1988). The pattern of gyrification in the cerebral cortex. *Anatomy and Embryology*, 179 (2), 173-179.

Insight: What Is It, Exactly?

A Commentary on Ursula Voss & Allan Hobson

Lana Kühle

In “What is the state-of-the-art on lucid dreaming? Recent advances and questions for future research”, Ursula Voss and Allan Hobson provide a detailed view of the features characterizing lucid dreaming and put forward four innovative hypotheses to explain why and how lucid dreaming occurs, as well as how lucid dream states are related to other states of consciousness. Their aim is to show that not only is there benefit to studying lucid dreaming in itself, as this would give us a deeper understanding of dream consciousness, but also that it is an important endeavor because of the kind of conscious state lucid dreaming is. To be sure, Voss and Hobson make important in-roads into the empirical study of lucid dreaming that ought to sprout new and exciting research in the area. As I will show, however, there remains much conceptual work to be done. In this commentary I tease out three aspects of Voss and Hobson’s view that would greatly benefit from philosophical consideration. First, I highlight the lingering confusion with what exactly insight is, and I point to how one might go about clarifying this notion. Second, I argue that our understanding of insight and meta-awareness in lucid dreaming could be greatly increased by looking at how these concepts are used and understood in relation to meditative states. Last, I explore the role of the body in lucid dreaming and argue that one’s bodily awareness in lucid dreams is far more multi-faceted than at it might at first seem.

Keywords

Bodily awareness | Consciousness | Dreaming | Insight | Lucidity | Meditation | Meta-awareness

Commentator

[Lana Kühle](#)

lkuhle@ilstu.edu

Illinois State University

Bloomington-Normal, IL, U.S.A.

Target Authors

[Ursula Voss](#)

voss@psych.uni-frankfurt.de

Johann Wolfgang Goethe-Universität

Frankfurt a. M., Germany

[Allan Hobson](#)

allan_hobson@hms.harvard.edu

Harvard Medical School

Brookline, MA, U.S.A.

Editors

[Thomas Metzinger](#)

metzinger@uni-mainz.de

Johannes Gutenberg-Universität

Mainz, Germany

[Jennifer M. Windt](#)

jennifer.windt@monash.edu

Monash University

Melbourne, Australia

1 Introduction

In “What is the state-of-the-art on lucid dreaming?—Recent advances and questions for future research”, [Ursula Voss & Allan Hobson \(this collection\)](#) aim to defend the veracity of, and value in empirically studying lucid dreaming. They provide a detailed view of the features characterizing lucid dreaming as well as hypotheses for why and how lucid dreaming occurs. As they claim, not only is there benefit to studying lucid dreaming in itself, as this would give us a deeper

understanding of dream consciousness, it is also an important endeavor because of the kind of conscious state lucid dreaming is. The authors argue that the study of lucid dreaming will also deepen our understanding of the structure of consciousness more broadly—the nature of meta-awareness, the notion of a self, and its relation to our ability to be meta-aware, etc.

To be sure, I think that Voss and Hobson make important in-roads in defending the vera-

city of lucid dreaming and putting forward hypotheses that ought to sprout new and exciting research in the area, as I will elaborate in section 2. However, I think there remains a need for caution in how we describe and define lucid dreaming, a great need for further clarification of what lucidity involves, and potentially fruitful connections to be drawn between lucid dreaming states and meditative states. In what follows, my goal is to elaborate on each of the following three points with a view to generating future discussion and discovery not only in the area of lucid dreaming research, but also in areas of meditation research and embodied awareness research.

The first point on which I focus—in section 3—is the concept of “insight”. To be sure, Voss and Hobson do offer us a definition of insight—an awareness of being in a dream, knowing that what one is currently experiencing is not real, etc.¹ However, their definition conflates and confuses whether the insight involved in lucid dreaming is a state or an ability, and whether it is an epistemic or phenomenal state/ability. In other words, does it involve knowledge of something, is it simply experiential, or is it an ability to do or know something, etc.? In this section, then, I delve deeper into what the authors mean by “insight” and explore these questions, as well as inquire whether insight is best understood using epistemological or phenomenological frameworks. Moreover, I consider what the consequences of an underdeveloped understanding of the concept of insight might be for the current state of research on lucid dreaming.

The second point on which I focus—in section 4—is the authors’ suggestion that we look at other states of waking consciousness with a view to determining how exactly insight comes to co-occur with REM sleep. I consider the potential similarities between lucid dreaming and meditation, and suggest that there are fruitful connections to be drawn between the meta-awareness associated with insight in lucid

dreaming and the meta-awareness involved in certain meditative practices.

The third point I consider—in section 5—is the experience of the body in lucid dreaming. In particular, I argue that if we accept one of the authors’ hypotheses—the Hybrid State Hypothesis—then we can enrich our understanding of the bodily awareness involved in lucid dreaming by looking at certain accounts of bodily awareness in waking consciousness. More specifically, I offer one interpretation for why the dual experience of the dream body and the real body in lucid dreaming is said to demand a lot of concentration by appealing to my recent work on bodily awareness in waking experiential consciousness. Before I begin exploring each of these three points, however, let me first summarize Voss and Hobson’s important contributions.

2 Voss & Hobson—A summary

In their piece, Voss and Hobson consider the latest empirical evidence on lucid dreaming and set forth four hypotheses that, they suggest, would begin to explain the whys and the hows of lucid dreaming. The four hypotheses proposed—the BMH (Brain Maturation Hypothesis), the GBH (Gamma Band Hypothesis), the HSH (Hybrid State Hypothesis), and the SCH (Space of Consciousness Hypothesis)—are based on five years of scientific research on lucid dreaming and, together, are meant to provide a multi-faceted picture of what lucid dreaming is, how it arises, why it arises, and how it relates to other states of consciousness.

The first hypothesis they propose is the BMH (Brain Maturation Hypothesis), which serves as a potential explanation for *why* there is lucid dreaming. Evidence shows that lucid dreaming occurs naturally and most often during certain periods of brain development and maturation in children and young adults.² The empirical evidence also suggests that lucid dreams are peculiar mental states that occur during the final stages on frontal lobe integration and, as such, are “nothing but an

¹ See Voss and Hobson’s target article in this collection, and their development of the LuCiD (Lucidity in Dreams) scale in Voss et al. (2013).

² See Schredl & Erlacher (2011), as well as the Voss & Hobson target article (this collection).

accidental confounding of conscious states during a time of high cerebral diversification” (Voss & Hobson [this collection](#), p. 8). For these reasons, Voss & Hobson hypothesize that “during childhood and puberty, frontal lobe activity is sometimes decoupled from the arousal state so that frontal lobes can become active in a state for which this type of activity is untypical”—the BMH ([this collection](#), p. 8). This, they propose, explains *why* lucid dreaming occurs.

Voss and Hobson then offer three other hypotheses—GBH, HSH, and SCH—as explanations of *how* lucid dreaming occurs. The GBH (Gamma Band Hypothesis) provides an account of how lucid dreaming arises by appealing to specific changes in brain activity associated with the onset of a lucid dream during ongoing REM sleep. Specifically, this hypothesis holds that the principle brain correlate of lucid dreaming is 40Hz activation of the frontal cortex—activation at this frequency brings about the meta-awareness associated with secondary consciousness. The HSH (Hybrid State Hypothesis) & SCH (Space of Consciousness Hypothesis) shift away from particular brain activity and, rather, provide a brain-based explanation and classification, respectively, of what lucid dreaming is in relation to other mental states. The HSH suggests that lucid dreaming involves elements of both waking and dreaming consciousness, and is, indeed, a destabilized hybrid state involving both frontal cortex activation, as suggested by the GBH, and REM sleep cortical activation. The HSH explains the *how* of lucid dreaming by offering a way to reconcile the subjective reports of lucid dreamers with the empirical data of cortical activation. The SCH lays out a three-dimensional model with which to categorize various states of consciousness and to see how the spectrum of mental states relate to one another along certain variables. This model allows us to situate lucid dreaming within a state space of consciousness and ascertain the similarities it might hold with other waking states of consciousness. These four hypotheses work together to consolidate the quantitative and qualitative data on lucid dreaming and provide a picture of why and how lucid dream-

ing occurs. For my purposes here, I will set aside the BMH and the GBH and will instead return to the HSH and the SCH in sections 4 and 5.

Importantly, the authors specify that their interest lies in considering REM-sleep lucid dreaming. In other words, the focus of their paper is to consider cases where the dreamer correctly achieves insight into the fact that he or she is dreaming while the dream continues (see Voss & Hobson [this collection](#), p. 4). The authors appeal to the Lucidity and Consciousness in Dreams Scale (LuCiD) they developed to assess the various features of a lucid dream state, and with this they describe eight features of lucid dream consciousness: insight, realism, control, memory, thought, positive emotion, negative emotion, and dissociation.³ Of these eight factors, three are highlighted as particularly important to the study of lucid dreaming—insight, control, and dissociation—as they do not typically appear in non-lucid dreams.⁴ The core criterion of lucid dreaming, however, appears to be insight. This feature, once it appears, then causally enables the possibility of control and dissociation. One of the issues that I will explore further in the next section is whether insight should be thought of as an epistemic or a phenomenal state, and what either of these interpretations might mean for understanding the role of insight in lucid dreaming.

Most of Voss and Hobson’s article discusses the features of insight and dissociation in relation to recent empirical evidence, and although there is indeed very illuminating discussion of these features, I nonetheless think there is still much conceptual confusion and semantic vagueness with regard to what exactly they are and how they relate to our non-dreaming conscious states. As I show in the next section, this is where philosophical considerations can help clarify the conceptual landscape and help move the empirical project forward.

³ Voss and Hobson don’t discuss the possibility of there being varying degrees of lucidity, and thus how these features might relate to such varying degrees. For a discussion of this, see Noreika et al. (2010).

⁴ There are rare cases where some of these aspects do occur in non-lucid dreaming states. See Voss et al. (2013) and Voss et al. (2014).

3 Understanding insight

The first element of Voss and Hobson's piece on which I want to focus my attention is the concept on insight. More specifically, I want to explore what the notions of lucidity and insight involve and how they relate to dream consciousness. As the authors clearly state throughout their paper, lucidity involves insight, and insight seems to be the key feature of lucid dreaming as it serves the basis of dream lucidity and enables the other elements of dream lucidity to arise, e.g., dissociation, control, etc. Without insight, it appears, one could not have lucid dreaming. Or, at the very least, it seems conceptually essential to have insight in order to be in a state of lucid dreaming.⁵ Given the importance of insight, it is key that we obtain a clear view of precisely what it is.

In the first place, I think it is necessary to distinguish between the *state* of insight and what one has insight about—let us refer to this as the *content* of insight. With regards to the state of insight, it is not so clear what this precisely is, and the authors do not adequately clarify it. For example, if it is an epistemic state, then it would have an intentional object. The questions then become: what are the intentional objects of the state of insight? What kind of knowledge does the state of insight involve? It is in the second section of their paper, "Quantification of Dream Lucidity as Subjective Experience", that Voss and Hobson attempt to describe and define what the state of insight is. There, they liken insight to a subjective awareness of our mental state. This subjective awareness, they go on to claim, is a form of secondary awareness, or meta-awareness that arises in lucid dreaming. They define meta-awareness, following Metzinger (2013), as "an instance of actively acquired self-knowledge or a sudden insight, regardless whether it is accurate or counterfactual" (Voss & Hobson this collection, p. 4). In short, insight appears to be a form of awareness that arises out of a more primary

awareness, and it allows the subject to attend to, or "see" what is occurring in primary awareness.

Now, a number of questions and issues arise from this definition of the state of insight. First, it seems quite problematic to define insight as a form of meta-awareness, and then to define meta-awareness as an instance of sudden insight. Perhaps, however, we might want to rely on the first half of the disjunct in the definition quoted above and understand insight as a form of actively-acquired self-knowledge. Given that the authors refer to insight as a form of reflection (Voss & Hobson this collection, p. 6) and as a form of knowing (ibid. p. 8) elsewhere in the text, I will assume that this is the more accurate reading of the definition. However, this still raises questions. In what way are we to understand "actively acquire" in the case of lucid dreaming? What does the dreamer *do* in a non-lucid dream state to acquire insight and thus bring about lucid dreaming? Is lucid dreaming an ability?⁶ If so, then perhaps it is trainable. Trainability might, in turn, provide us with an answer to the first two questions: namely, what might be involved in actively acquiring insight and what exactly the dreamer does. If it is an ability, perhaps the ability in question is one of moving into a state of meta-awareness. Moreover, if the ability to shift into a state of meta-awareness is an element of what the subject "does" to actively acquire insight while dreaming, then looking to other mental states that involve meta-awareness and that are also "trainable" could be beneficial.

One such set of mental states that involve an aspect of trainability are meditative states. Meditation is a practice, and with practice one is able to achieve and sustain certain forms of awareness—focused attention, open awareness, etc.⁷ If we take the element of practice in meditation as being akin to a form of trainability, and the forms of awareness in meditation to be

⁶ For a review of the ways in which lucid dreaming is trainable see Stumbrys et al. (2012).

⁷ Focused attention meditation involves developing one's ability to concentrate on an object for an unlimited amount of time. Open presence/awareness meditation involves opening one's awareness to all experiential aspects of the moment, e.g., mental states, bodily sensations, environmental stimuli, etc., and not attending to anything in particular.

⁵ We might not, however, be warranted to make a similar empirical claim, i.e., that insight is empirically essential *and* sufficient for lucid dreaming. Indeed, there is controversy over whether insight is empirically sufficient for lucid dreaming. See Voss et al. (2013) and Windt & Metzinger (2007) for further discussion of this issue.

akin to meta-awareness, then looking at the practice of meditation—what one does, how one improves, and so on—might be informative in ascertaining whether actively acquiring insight in lucid dreaming is something that is trainable.⁸ As I will detail in the next section, I believe there are also other reasons to consider meditation in relation to lucid dreaming.

Another line of questioning that arises from Voss and Hobson's definition of the state of insight relates to the concept of self-knowledge that, they claim, is an element of insight. How are we to understand the concept of "self-knowledge" as it applies to the insight gained in lucid dreaming? What is the "self" involved? And how strict a use are we making of the concept of knowledge—do we mean a justified true belief? The state of insight seems to involve very different characteristics. Voss and Hobson hold that insight involves knowledge, or the realization that one is dreaming, and they also describe insight as an experiential phenomenon, and one that involves reflection. The issue here is that "knowledge", "realization", "experiential phenomenon", and "reflection" are not interchangeable concepts. It remains quite unclear from the descriptions of insight provided whether we should view the state of insight as an epistemic or phenomenal state of consciousness. Based on the information Voss and Hobson provide in their piece, I am inclined to move away from an epistemological view of the state of insight as I think the concept of self-knowledge is too complex for the phenomenon that Voss and Hobson describe. What I mean here is simply that with the concept of self-knowledge come notions of identification, veridicality, the self, and so on, and I do not think that such a complex concept is necessary to account for the *experience* of insight in lucid dreaming. As Voss & Hobson explain, insight is "[t]o some extent, the dreamer [having]"—"however limited"—"access to secondary consciousness, enabling her to reflect on her present state" (this collection, p. 8), and "[b]y secondary consciousness we mean the subjective

awareness of our state in dreaming" (*ibid.*, p. 4). Instead, I would suggest using the concept of self-awareness to capture what is involved in insight, and by self-awareness I mean here simply the awareness of being in a certain experiential moment.⁹ So, in the case of insight, one becomes aware of dreaming—a self-awareness—rather than acquiring the self-knowledge that one is dreaming. Perhaps, however, there is reason to separate the concept of insight from that of lucidity, and with this distinction we might want to describe lucidity as a phenomenal state and insight as an epistemic state. I think there might be good reason to take this route, and I explore this in the next section by considering the potential relation between insight in lucid dreaming and insight in meditative states.

Now, these are issues that arise when considering what is meant by the "state" of insight. As I distinguished earlier, however, there is also the "content" of insight. With regards to the content of insight, in cases of lucid dreaming things are relatively clear: one gains insight on the nature of one's current dream state, i.e., that one is currently dreaming. In other words, insight involves coming to realize *that* one is dreaming. This way of describing what occurs in insight, however, could be seen as problematic in that it takes insight to involve a particular kind of knowledge, namely, knowledge-THAT. If indeed insight involves knowledge-THAT, then this opens the door to theory-contamination; that is, the content of insight is contaminated by what one already believes about dreams, consciousness, etc.¹⁰ Although I grant that this issue shows that there is a need to clarify what exactly the content of insight is, I am uncertain that it is as problematic as it might at first seem to hold that insight involves knowledge-THAT. How else would one be able to "realize" that one was dreaming if one was not able to identify, to some degree, that the state one is in is a dream state? Moreover, it certainly seems that to perform such an identification one

⁸ The Tibetan Buddhist practice of dream yoga is a particularly interesting area worthy of exploration in relation to this issue. See LaBerge (2003) for a discussion of dream yoga in relation to lucid dreaming research.

⁹ The "self" in self-awareness here does not refer to an ego or any robust notion of a self. Moreover, the kind of awareness I'm suggesting is not a categorical awareness, i.e., an awareness of the experiential moment as belonging to a category of consciousness (see Metzinger 2009). Rather, it is meant simply to point to a reflexivity of awareness (see the concept of "pre-reflective self-awareness" in Zahavi 2005).

¹⁰ Thanks to Thomas Metzinger for pointing out this issue.

would rely on theory-contaminated beliefs—certain conceptions of what a dream is like, etc. Perhaps there is no way of avoiding theory-contamination altogether, and thus the issue becomes one of determining how much contamination is allowable in the case of insight.

I certainly grant that given the state of research into lucid dreaming—it is still very much in its infancy, no doubt—it is not unexpected that a clear understanding of a complex concept such as “insight” is still lacking. To be sure, the authors have provided a good starting point for developing a full description of the state of insight. However, given that it is, arguably, the key element of dream lucidity, I worry about how well we can empirically investigate, or interpret our empirical findings of the whys and hows of lucid dreaming if we don’t first ensure that we have a working understanding of insight. To define insight as a form of meta-awareness, or secondary consciousness that involves actively acquired self-knowledge, is not informative enough to allow us an understanding of what insight in dream consciousness is or why it is so special and important.

To be sure, I think it would be entirely inappropriate to hold Voss and Hobson accountable for not teasing out the concept of insight further. They are empirical researchers, and as such have paved the way for future research in this area. However, I think that the lack of conceptual clarity and the semantic vagueness that remains in this area point to the need for philosophical inquiry and the value of integrating philosophical work with empirical work on lucid dreaming. It now lies in the hands of philosophers to ensure that the future progress of this research is based on a strong conceptual foundation. One direction to take in this endeavor is to follow Voss and Hobson’s suggestion and look at other areas of research concerned with meta-awareness, reflection, and insight. In the next section, I propose that one such area is that of meditation.

4 Lucidity, meta-awareness, and meditation

The second point I want to focus on is Voss and Hobson’s desire to consider other states of con-

sciousness to better understand the state of lucid dreaming. In particular, they express an interest in considering altered states such as hypnosis or mind wandering. I suggest that there might also be benefit in considering meditation. Specifically, I think we can fruitfully make use of how the notion of insight in meditative experiences is developed to clarify that of insight in lucid dreaming. We would first have to show that there are enough important similarities between the notion of insight involved in meditation and the notion of insight involved in lucid dreaming, and this will be my aim in what follows.

To be sure, there are many and various meditation styles and practices, each with its own experiential path to higher states of awareness. Broadly speaking, there are three categories of meditative practice, each with variants, and there is overlap in some respects between the categories.¹¹ First, there is focused attention meditation—this involves developing one’s ability to concentrate on an object for an unlimited amount of time. Second, there is open presence meditation—this involves opening one’s awareness to all experiential aspects of the moment, e.g., mental states, bodily sensations, environmental stimuli, etc., and not attending to anything in particular. Third, there is insight meditation—this involves developing mindfulness or meta-awareness over one’s mental states. More specifically, and most interestingly when compared to the concept of insight in lucid dreaming, “[insight meditation] is also one of the earliest and most fundamental forms of meditation. For Buddhist theorists, [insight meditation] is a style of meditation that, in combination with the focus or stability provided by cultivating [focused attention], enables the practitioner to gain insight into one’s habits and assumptions about identity and emotions” (Lutz et al. 2007, p. 504). For my purposes here, I will set the finer variations among these three main styles of meditation aside since I’m merely concerned with drawing out the similarities, in broad strokes, between the sought-after meditative state and the insight it is intended to provide,

¹¹ See Lutz et al. (2007) for a more detailed account of the various styles of meditative practice and their historical roots.

and the lucid dreaming state and the insight required to bring it about. Interestingly, however, the concept of insight applied to the practice of insight meditation is quite similar in many respects to the concept of insight applied to the experience of lucid dreaming.

To be sure, the concept of insight, as it relates to meditation, is very complex, and also not fully defined. There are many levels of insight, and many aspects of mental life, the self, and life more broadly that one achieves insight about, depending on the style of meditation one engages in and the level of mastery one develops in one's meditative practice. For example, in the practice of focused attention meditation, a novice practitioner might be said to have gained insight upon becoming aware of the difficulty involved in maintaining attention on the flow of the breath through the nostrils. The insight here is of a particular aspect of mental life, namely, the fleeting nature of attention. Whereas in the case of an experienced practitioner with hours of meditative experience, the insight gained may involve the nature of the self—for example, that it is characterized by desire and craving, or that it is ultimately an illusion. Nevertheless, I think that we can certainly make use of the way the concept of insight is broadly understood in meditation to clarify its relation to lucid dreaming, if it has any relation.

First, I take it that when we speak of insight gained through meditation, we aren't referring to a particular state that is achieved, but rather to a form of knowledge that is gained within a state of consciousness. The state from within which we might be said to achieve insight is a state of meta-awareness, but being in this state doesn't necessarily imply that insight has been achieved. For example, the novice practitioner may become meta-aware of what it is like to try to maintain focused attention on the breath, but this doesn't necessarily mean that he gains knowledge from this about the nature of attention and consciousness more broadly. Conversely, it seems that in the case of lucid dreaming, at least as described by Voss and Hobson, insight is understood to be synonymous with meta-awareness. This seems a natural understanding given that, as per Voss

and Hobson, when lucidity is achieved there is necessarily insight. That is, one could not, it appears, be meta-aware of their dreaming without having insight into the fact that they are dreaming. However, is this really *insight*? This is where I think we may want to tease apart the notions of lucidity and insight, following our understanding of meta-awareness and insight in cases of meditation.

In the case of lucid dreaming, there certainly is the experience of coming to realize one is in a dream state. This is the phenomenological interpretation of the state of insight I discussed in the previous section—what I also called the self-awareness of dreaming. However, we may want to refer to this aspect of lucid dreaming as lucidity, rather than insight. In other words, when lucidity occurs while dreaming, why should we not be satisfied saying that one has simply become aware of their dreaming? Why should we take this to be insightful? Maybe because lucidity doesn't merely involve a passive awareness of the dream state, but also an understanding by the dreamer of *what* she has become aware of—and this enables dissociation, plot control, etc. The suggestion that there is now an understanding that the dreamer has of being in a dream, however, brings into the picture the epistemological interpretation mentioned earlier. Given this, insight is better viewed as an epistemic state. In fact, maybe there is not only a need to dissociate lucidity from insight in the case of lucid dreaming; we may want to grant that both admit to phenomenological and epistemological degrees.¹² As we see in meditation, there are many levels of insight—many areas of our existence of which we can gain knowledge—and so maybe there is also reason to think that there are further forms of insight to be had in lucid dreaming as well. One particularly interesting point of convergence between the empirical work on lucid dreaming and meditation is in the phenomenon of dream yoga.¹³ As a result, we might not want

¹² This very idea has been explored in Windt & Metzinger (2007), as well as in Noreika et al. (2010).

¹³ In particular, the case of Tibetan dream yoga mentioned earlier, which involves using meditative practice in the dream, might be an instance of exploring just how meditation and lucid dreaming can come together, and could be informative for our understanding of

to define insight as a state of consciousness, or as a meta-awareness. Rather, we may instead see insight as a form of knowledge that accompanies lucidity, and lucidity as a form of meta-awareness.

Another area of similarity between meditation and lucid dreaming that I want to explore lies in the structure of each of these experiences.¹⁴ Both seem to involve some form of dissociation. As [Voss & Hobson \(this collection\)](#) describe, “lucid dreams can be considered dissociated states of consciousness in which the dream Self separates from the ongoing flow of mental imagery. The dream is still a dream but the person is able to distance him/herself from the ongoing imagery and may even be successful in gaining (at least partial) control over the dream plot” (pp. 8–9). The experiential feature of separation of the dreamer from the dream while the dream continues to unfold is akin to the observational stance that one strives to take in meditation, in particular in focused meditation. When meditating, one aims to become aware of one’s stream of consciousness—one tries to separate oneself, as it were, from the stream of thoughts, beliefs, desires, etc., in order to become aware of its transient nature. For example, one becomes aware of, say, the fleeting nature of attention and mental life. Similarly in lucid dreaming, one becomes aware of being in a dreaming state.

However, the concept of “self” that seems to underlie Voss and Hobson’s discussion of lucid dreaming is quite different from how the self is understood in meditation. Voss and Hobson appear to have a very robust sense of self at play, and I’m not quite sure why this is so, or whether we want to bring such a conception of self into the picture. One of the most telling passages in their article, and one that I find most problematic is the following:

both the nature of meditative states and that of lucid dreams. As [LaBerge](#) notes, “for more than a thousand years Tibetan Buddhists have believed that it is possible to maintain the functional equivalence of full waking consciousness during sleep. This belief is not based on anything as tenuous as theoretical grounds but upon firsthand experience with a sophisticated set of lucid dreaming techniques collectively known as the Doctrine of Dreams or dream yoga” (2003, p. 233).

¹⁴ See [Evan Thompson’s](#) entry in [this collection](#), as well as [Thompson \(2014\)](#).

This fits well with the common description of lucid dreams as (partial) awakening in your dreams and of involving a split between dreamer and dream observer who coexist and change relative dominance of the mind at will ([Occhionero et al. 2005](#)). The implications of this line of reasoning have profound impact on the theory of mind. There are two selves suggesting that the self is a construct elaborated by the brain ([Metzinger, 2003, 2009, 2013a](#)). The two selves of the lucid dreamer [...] ([Voss & Hobson this collection](#), p. 9, emphasis added).

Why would we want to describe the result of the dissociation in lucid dreaming as one that involves a split between a dreamer self and a dream-observer self? Furthermore, on the basis of what would there be reason to argue that the self is a construct?

If the experience in lucid dreaming is one of shifting back and forth between being meta-aware of being in dream consciousness and being in the dream itself as the dreamer, why would we not want to speak of this as a change in experiential perspective rather than as an experience of two selves?¹⁵ Moreover, if we look to how similar meditative experiences are described, we don’t speak of there being two selves, the self within the stream of consciousness and the self that observes the stream of consciousness. Rather, we speak of our shifting experiential perspectives wherein we move, as a single subject of experience, from being within the flow of consciousness to observing the flow of consciousness. Furthermore, one of the insights gained from meditative practice is that there is indeed no self.

I grant that it is perhaps in keeping with the subjective reports of lucid dreamers to speak of two selves in the lucid dream state. If the subjective report that [Voss & Hobson](#) quote in their paper ([this collection](#), p. 9) is but one example of the way in which subjects describe their experiences, then it certainly seems nat-

¹⁵ The shift in experiential perspective might even be more complex than this; see [Rosen & Sutton \(2013\)](#) for an interesting discussion of self-representation in dreams.

ural to take on such a view of the self. However, I suspect that the subjective reports may be constructed in a manner that is biased by a certain colloquial manner of speaking about the self,¹⁶ and thus don't rightly capture if and what the self is in relation to the structure of consciousness. Certainly I am not suggesting that we shouldn't take the subjective reports seriously—indeed I think that they provide invaluable information into the phenomenology of lucid dreaming. However, we must be careful to properly interpret these reports, and perhaps this will involve developing ways to discover whether certain biases have come into play in the subject's report of her experience, and how these biases have affected the qualitative data.

5 The hybrid state hypothesis and bodily awareness

The third and last point I want to consider is the place of the body, and bodily awareness, in lucid dreaming. I was particularly struck by two lucid dreamer reports. The first is the one that [Voss & Hobson](#) quote in their paper wherein the lucid dreamer explains that “[i]n these short periods of lucidity the awareness of the acting dream body and the real body in bed exist simultaneously and it costs a lot of concentration to keep the balance between both” ([this collection](#), p. 9). The second comes from Dutch psychiatrist Frederik van Eeden, who coined the phrase “lucid dreaming”:

In January, 1898 [...] I was able to repeat the observation. [...] I dreamt that I was lying in the garden before the windows of my study, and saw the eyes of my dog through the glass pane. I was lying on my chest and observing the dog very keenly. At the same time, however, I knew with perfect certainty that I was dreaming and lying on my back in my bed. And then I resolved to wake up slowly and carefully and observe how my sensation of lying on my chest would change to the sensation of lying on my back. And so I did, slowly and

deliberately, and the transition—which I have since undergone many times—is most wonderful. It is like the feeling of slipping from one body into another, and there is distinctly a double recollection of the two bodies. I remembered what I felt in my dream, lying on my chest; but returning into the day-life, I remembered also that my physical body had been quietly lying on its back all the while. This observation of a double memory I have had many times since. It is so indubitable that it leads almost unavoidably to the conception of a dream-body. ([van Eeden 1913](#))¹⁷

I found the description of there being two bodies rather interesting, and, particularly in the subject report cited by Voss and Hobson, the mention of the cost of concentration to be very intriguing. To be sure, there is but one physical body, namely the one lying in bed. Yet the dreamer experiences both the body in bed and the body with which she is engaged in the dream, and finds it somewhat demanding to maintain an experiential balance between both. In this last section, I put forward an explanation of this experience by relying on the Hybrid State Hypothesis alongside my work on bodily awareness during waking consciousness.

According to the HSH Voss and Hobson put forward, lucid dreaming is a hybrid state with both elements of waking and dream consciousness. This is so because there is a dissociation that occurs between the dream self and the ongoing dream imagery. Physiologically, although brain activity associated with REM sleep continues, in lucid dreaming there arises, in addition, brain activity in parts of the brain associated with conscious awareness and executive ego functions. The hypothesis, then, is that “lucid dreams push the arousal system towards waking yet remaining within the region occupied by REM sleep [...]. Lucid dreaming is, thus, a fragile, destabilized hybrid state” ([Voss & Hobson this collection](#), p. 9). If this hypothesis is correct, then there may be value in looking at how we are aware of our body in a waking con-

¹⁶ This, as Metzinger would point out, would be another instance of theory contamination.

¹⁷ Thanks to Metzinger for pointing out this classical description of a lucid dream experience.

scious state to help better understand the seeming duality of bodily awareness involved in lucid dreams. More specifically, if we take seriously the above-quoted subjective report, then the hybrid state hypothesis in combination with certain hypotheses about bodily awareness in waking conscious states might shed light on how the experience arises.

What I find particularly interesting about the reports are two things:

- a. the simultaneous experience of a dream body and the real body in bed; and
- b. the amount of concentration needed to keep the balance between both.

In regards to the first, I find myself wondering the following: what does the subject mean by simultaneous, here? Does she mean that both bodies are experienced *at the same time*, or rather, that there is a very quick and continuous shift back and forth from the dream body to the real body, such that it *seems* like they are both being experienced simultaneously? I am inclined to think that what is happening is a very quick attentional shift back and forth between the two “bodies”. My reasons for thinking this come from how I account for our bodily awareness in waking life.

I take it that in our everyday experiential lives we are aware of our body both as an object and as a subject. The distinction between awareness of the body as object and as subject stems from the Phenomenological tradition¹⁸ and it is best understood as follows. I can be said to be aware of my body as object when I direct my attention to my body and thereby perceive it as I would any other object in the world. The key characteristic of our awareness of the body as object is that it is attentional. Alternatively, I can be said to be aware of my body as subject when I am aware of my body as that *through* which I experience the world—not as an object onto which I turn my attention, but rather as that which engages with my environment. My awareness of my body as subject is

also referred to as a bodily self-awareness, and it is characterized by an inattentional awareness—a form of awareness that does not involve holding attention to an object.¹⁹

Now, my typical experiential consciousness involves a bodily self-awareness, although it doesn’t always involve an awareness of the body as object. This is because I don’t always attend to my body. Take, for example, my sitting in a chair reading a book. Typically, my attention lies with the book—I focus on the words on the page, say. In attending to the book, I don’t simultaneously attend to my hands holding the book, although they are certainly a part of my overall experience insofar as they don’t disappear from my awareness entirely. I certainly can shift my attention to my hands, and thereby become aware of them as object; however, in doing so, I contend, I am no longer attentively aware of the words I was reading a moment ago. In fact, I take it that if I were to try to be aware of my hands and the words on the page simultaneously, I would find this quite difficult as it would involve a continuous and rapid shift in attention back and forth between the words and my hands. I think a similar account holds in the case of lucid dreaming with regard to the dream body and the real body.

I propose that in the case of one’s bodily awareness in lucid dreaming, the real body is experienced both as subject and as object. It is the subject’s actual body, and therefore one that she is aware of as subject, but in addition her experience of her real body, in the lucid dream, is of her body as an object—she becomes aware of her body as object by her attention shifting to it momentarily. However, her attention does not remain with her real body; instead it quickly shifts back to the dream body as well. In that experiential moment, the dream body becomes an object for her as she attends to it. I think the further clue as to why we should interpret the experience of the body in lucid dreams as one of shifting attention, and even perhaps competing attention between the real and the dream body, comes from the second element of the subject’s report men-

¹⁸ A philosophical tradition most often associated with the work of Husserl, Merleau-Ponty, Sartre, etc.

¹⁹ I develop this distinction further in my thesis, “Embodiment and Subjectivity—the Origins of Bodily Self-Awareness”.

tioned above—the claim that “it costs a lot of concentration to keep the balance between both”.

Why is there a need to keep a balance between the real and the dream body? Perhaps because, as the HSH suggests, there are elements of both waking and dreaming states at play. If we take bodily awareness to be a fundamental element of waking consciousness—or even consciousness tout court, as I do—as well as a key element of dream consciousness, then it makes perfect sense that in a lucid dream the subject finds herself with these two bodies that must be balanced in the same way that the waking and the dream states must be balanced to remain in the lucid dreaming state.²⁰

The question then becomes: why does it cost a lot of concentration to maintain this balance? I think the answer to this question brings us right back to my suggestion above, namely that the simultaneity of the dream and real body experience is one of shifting, or even competing attention. If there is a continuous shift in attention, rather than a joint experience of both bodies, then this would explain the apparent cost of trying to maintain concentration on both bodies in a lucid dream state. It would be like walking a tightrope, trying to avoid leaning too far to the right or too far to the left, and doing so by continuously shifting your body to maintain that balance. It would require an incredible amount of concentration—in a general sense, one experiences everything all at once, but in a more precise sense, one’s attention is continuously shifting between one’s body and one’s environment in order to maintain balance.²¹

One last point of inquiry. As I mentioned above, there is a distinction to be made in accounting for our bodily awareness in waking experiential consciousness between our awareness of the body as object and our awareness of the body as subject, i.e., bodily self-awareness.

²⁰ The place and role of the body, and our bodily awareness in lucid dream states, is far more complex than I can show here—in fact, there are instances of bodiless dreams. Although a complete consideration of these issues is beyond the scope of this commentary, an excellent discussion of this topic can be found in Windt (2010).

²¹ This is also how lucid dreams are commonly described in the literature, i.e., as a balancing act. See LaBerge (1985) and Brooks & Vogel-song (2000).

However, I wonder if a similar distinction might also apply in cases of lucid dreaming given the HSH. In other words, is there a bodily self-awareness—of the real body or even the dream body in a lucid dreaming state? And, if so, how does it relate to the awareness of the dream body and the real body described by subjective reports? To begin answering these questions we would need to explore the subjective reports of lucid dream experience in relation to bodily awareness more specifically. Perhaps we might begin by looking back upon the report by van Eden. Indeed, I certainly take this to be an interesting avenue of exploration given the ever-increasing interest in taking an embodied approach to consciousness.

6 Conclusion

In closing, let me review the three points of inquiry on which I chose to focus here. First, I inquired as to what exactly the concept of insight involves in the case of lucid dreaming and whether we should think of insight as a phenomenal or epistemic state. I suggested that the lack of clarity with regard to the concept of insight shows the need for rigorous philosophical inquiry with a view to laying down a solid conceptual foundation from which to pursue future empirical research. Second, I inquired as to how meditation and lucid dreaming are similar and where research on meditation might provide information to research on lucid dreaming. I highlighted some interesting overlaps in the concepts of insight in meditative practice and lucid dreaming, and explored the feature of dissociation in lucid dreaming in relation to the notion of a self. Third, I looked at how we are aware of our body in lucid dreaming and considered whether our accounts of bodily awareness in waking consciousness can be used to inform our understanding of bodily awareness in lucid dreaming. I also suggested that the distinction between awareness of the body as object and of the body as subject used to describe waking bodily awareness could help us tease out the ways in which the body is experienced in lucid dreams.

As I stated above, the empirical study of lucid dreaming is still very new and, thus, still very much in an exploratory phase. As a result, it is easy to point out various areas for further inquiry and suggest avenues of future investigation. However, it is nonetheless important to acknowledge the work that Voss and Hobson have done to advance our understanding of the phenomenon of lucid dreaming. Not only have they provided a convincing account of why lucid dreaming occurs (BMH), they also put forward an interesting hypothesis for the neural basis of lucid dreaming (GBH). Moreover, their HSH and SCH will serve to further the conceptual analysis of lucid dreaming and its relation to other mental states across the spectrum of sleeping to waking consciousness. In short, I agree with Voss & Hobson that “the experimental study of lucid dreaming is a powerful paradigm for understanding the brain basis of conscious experience” (this collection, p. 4). Moving forward, we must now expand the area of research to allow for important philosophical considerations that will strengthen the conceptual framework underlying this exciting new paradigm.

References

- Brooks, J. E. & Vogelsong, J. (2000). *The conscious exploration of dreaming: Discovering how we create and control our dreams*. Bloomington, IN: AuthorHouse.
- LaBerge, S. (1985). *Lucid dreaming*. New York, NY: Ballantine Books.
- (2003). Lucid dreaming and the yoga of the dream state: A psychological perspective. In B. A. Wallace (Ed.) *Buddhism and science: Breaking new ground* (pp. 233-258). New York, NY: Columbia University Press.
- Lutz, A., Dunne, J. D. & Davidson, R. J. (2007). Meditation and the neuroscience of consciousness: An introduction. In P. D. Zelazo, M. Moscovitch & E. Thompson (Eds.) *The Cambridge Handbook of Consciousness* (pp. 499-551). New York, NY: Columbia University Press.
- Metzinger, T. (2009). *The ego tunnel: The science of the mind and the myth of the self*. New York, NY: Basic Books.
- (2013). The myth of cognitive agency: Subpersonal thinking as cyclically recurring loss of mental autonomy. *Frontiers in Psychology*, 4 (931). [10.3389/fpsyg.2013.00931](https://doi.org/10.3389/fpsyg.2013.00931)
- Noreika, V., Windt, J. M., Lenggenhager, B. & Karina, A. A. (2010). New perspectives for the study of lucid dreaming: From brain stimulation to philosophical theories of self-consciousness. *International Journal of Dream Research*, 3 (1), 36-45.
- Rosen, M. & Sutton, J. (2013). Self-representation and perspectives in dreams. *Philosophy Compass*, 8 (11), 1041-1053. [10.1111/phc3.12082](https://doi.org/10.1111/phc3.12082)
- Schredl, M. & Erlacher, D. (2011). Lucid dreaming frequency in a representative German sample. *Perceptual and Motor Skills*, 112, 104-108. [10.2466/09.PMS.112.1.104-108](https://doi.org/10.2466/09.PMS.112.1.104-108)
- Stumbrys, T., Erlacher, D., Schädlich, M. & Schredl, M. (2012). Consciousness and cognition. *Induction of lucid dreams: A systematic review of evidence*, 21 (3), 1456-1475.
- Thompson, E. (2014). *Waking, dreaming, being: Self and consciousness in neuroscience, meditation, and philosophy*. New York, NY: Columbia University Press.
- (2015). Dreamless sleep, the embodied mind, and consciousness. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- van Eeden, F. (1913). A study of dreams. *Proceedings of the Society for Psychical Research*, 26, 431-461.
- Voss, U., Schermelleh-Engel, K., Windt, J., Frenzel, C. & Hobson, J. A. (2013). Measuring consciousness in dreams: The lucidity and consciousness in dreams scale. *Consciousness and Cognition*, 22, 8-21.

- Voss, U., Holzmann, R., Hobson, A., Paulus, W., Koppehele-Gossel, J., Klimke, A. & Nitsche, M. A. (2014). Induction of self awareness in dreams through frontal low current stimulation of gamma activity. *Nature Neuroscience*, 17 (6), 810-812. [10.1038/nn.3719](https://doi.org/10.1038/nn.3719)
- Voss, U. & Hobson, A. (2015). What is the state-of-the-art on lucid dreaming? - Recent advances and questions for future research. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Windt, J. M. (2010). The immersive spatiotemporal hallucination model of dreaming. *Phenomenology and Cognitive Sciences*, 9, 295-326. [10.1007/s11097-010-9163-1](https://doi.org/10.1007/s11097-010-9163-1)
- Windt, J. M. & Metzinger, T. (2007). The philosophy of dreaming and self-consciousness: What happens to the experiential subject during the dream state? In D. Barrett & P. McNamara (Eds.) *The new science of dreaming, vol. 3: Cultural and theoretical perspectives* (pp. 193-247). Westport, CT: Praeger Perspectives/Greenwood Press.
- Zahavi, D. (2005). *Subjectivity and selfhood: Investigating the first-person perspective*. Cambridge, MA: MIT Press.

Reflections on Insight

A Reply to Lana Kühle

Ursula Voss

In this reply to Kühle, I will respond to her comments on the role of insight in lucid dreaming, especially regarding the question of whether it may be knowledge-based or instead express a sensorial experience. My answer rests on experimental findings, acknowledging Kühle's remarks, and taking her methodological challenges into account. I will challenge her proposal that insight might be called a state, opting for a definition of a transient thought atypically embedded within the state of dreaming, which may suffice to retrospectively call a REM dream lucid, but which will not satisfy the assumptions underlying the existence of a state.

Keywords

Insight | Lucid dreaming | Lucid scale | REM sleep

Author

Ursula Voss

voss@psych.uni-frankfurt.de

Johann Wolfgang Goethe-Universität
Frankfurt am Main, Germany

Commentator

Lana Kühle

lkuhle@ilstu.edu

Illinois State University
Bloomington-Normal, Illinois, USA

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

The commentary by Kühle reminds me of a remark made by a distinguished and renowned Swiss sleep researcher who asked me recently, during a lengthy discussion of our work on lucid dreaming, “how can you be sure that what you call a dream really exists”. In other words, he wanted to know how we could prove that dream narratives were memories of REM-sleep mental activity instead of, say, fantasies occurring during the process of awakening or memories of hypnagogic hallucinations, etc. It struck me then that I had neglected to openly postulate the key assumption that our work rested upon, namely that dreams really exist. So I still owe

him a detailed response and Kühle's commentary provides me now with the opportunity to generate an adequate reply. In the following, I will focus on Kühle's main argument, which seems to circle around the definition of “insight” and the question of whether it represents an epistemological statement or a phenomenological experience. I will shortly enter into discussion of whether it is justified to define insight as a state, as this assumption is not to be deduced from our work but certainly points to a need for clarification. While interesting, I will refrain from commenting on her speculations on whether insight may or may not be an ability

except for proclaiming that in my view, insight represents nothing but a result of neurobiological processes we still know far too little about. However, it is a fact that entering the state of lucid dreaming can be trained. Can insight *per se* be trained? I doubt it. Can the ability to generate insight be trained? According to recent studies on gamma-band activity in the developing and mature brain (see references in the main text), it is at least a possibility.

2 The role of insight in lucid dreaming

In her commentary, Kühle claims that the way we use the term “insight” leaves many—mostly philosophical—questions unanswered. While I certainly agree in principle that solving one question often generates many others, I also believe that there is some need for clarification regarding terminology. It seems that the discussion of what insight is and what it isn’t reveals one of the key methodological differences between our disciplines. Whereas philosophy of mind is mainly involved in meta-theory and the conceptualization of psychological theories, the focus of experimental psychology lies on the testing of hypotheses, albeit neither foci apply exclusively. By definition, however, experimental psychology aims at identifying cause-and-effect relationships between observable phenomena by applying experimental methods to induce controlled manipulations of so-called “independent variables”, leading to reproducible changes in “dependent variables”. Although experiments are hypothesis-based, testing specific (confirmatory) or unspecific predictions (exploratory) derived from theory, progress is often made when such an experiment leads to an unpredicted result. Such was the case in the construction of our LuCiD scale.

In the set of lucid and non-lucid dreams investigated and reported on by our group (Voss et al. 2013), we identified a factorial structure in which eight item clusters (which differed from the theoretically predicted ones) showed sufficient common variability to consider the items within each cluster related. These eight factors accounted for a large portion of variance in dream consciousness as

defined a priori, and based on theoretical considerations. The items in the item pool statistically identified as the single factor we referred to as “insight” pertained to the verbal communication that one *knew* one was dreaming while the dream continued. As such, insight would have to be regarded (in an epistemological sense) as understanding that at a particular moment within the dream, the dreamer acquired knowledge about his or her state of consciousness, which would be the hybrid state of lucid dreaming.

As Kühle correctly points out, this may or may not be true, however. It is just as possible that a dreamer who states upon his or her awakening: “I *knew* it was a dream while the dream continued” only thought that he or she knew, while in truth, he or she may have *sensed, felt, or experienced* that the ongoing dream action was not real. This would then pertain to a phenomenological experience similar to what Duncker (1947) refers to as “conscious participation” (p. 505), describing the sensorial experience that one is, at a particular moment, consciously aware of (pp. 508–509). On the other hand, even if we really experienced insight in a phenomenological sense, how can we be sure that this experience was not the result of the epistemological recognition of some sort of incongruence within the dream at some particular point in time? To me, this line of thought resembles that revolving around the question of whether we can be certain that a dream is really a dream and not something else. Philosophically, this is of course fascinating. But to experimental psychologists, such a discussion is unsettling because it is so difficult to translate into testable, i.e., operationalizable, hypotheses. Our admittedly very pragmatic approach is to define underlying assumptions such as “*we assume that dream reports generated from REM sleep awakenings are mentations generated during REM sleep and (fractionally) remembered (at least) until questioning*” or “*we assume that verbal accounts are reliable and valid*”. These assumptions can then again be challenged by separate experimental studies. In the case of doubting the existence of REM sleep dreams, an experiment

could be set up, for example, interrupting different states of arousal such as meditation, daydreaming, NREM sleep, or REM sleep and questioning the subject with respect to immediate recollections of mental activity. A comparison would lead to the conclusion that reports from REM sleep awakenings differ fundamentally from reports gathered from other states of arousal. This has, of course, been successfully achieved and repeated many times. However, the question is still not solved. It is doubtful, for example, whether an arousal from REM sleep enables as accurate a report as an arousal from the meditative state. Similarly, we cannot exclude the possibility that REM sleep alters mnemonic processes in a different way to NREM sleep, so that obvious discrepancies in NREM and REM reports are due to state-dependent retrieval and filtering processes and not at all related to different fantasies generated during the particular state.

In the same way, it certainly is appropriate to wonder about the true nature of what we refer to as “insight”. To psychologists, the explanation that a factor name is really only an attempt to describe a commonality between different but related observations is probably satisfactory. To philosophers, this will of course not be the case. However, with psychological pragmatism in mind, I would like to point to some empirical findings (and their immanent difficulties) regarding the question on how to further explore the nature of insight in lucid dreams: when we constructed the LuCiD scale (Voss et al. 2013), we started out with a set of 50 items that were selected on the basis of theoretical consideration. In a first step, these items were tested on a large sample of dreamers, leading to 158 dream narratives considered valid. These were then analyzed for factorial structure as well as for item reliability. Several items that might have been potentially informative regarding the question of epistemology vs. phenomenology proved either indistinct in differentiating between lucid and non-lucid dreams or they yielded too high statistical item difficulties so that they had to be eliminated from further evaluation. Some examples are:

- While dreaming my sensations were the same as when I imagine something or daydream during wakefulness
- While dreaming I was convinced that I was awake.
- I wasn’t in the dream, I had no self.
- While dreaming I felt that I knew where I was sleeping.
- While dreaming I was more than one person.

This finding of no-difference is of course by no means sufficiently informative to consider the question of insight in dreaming solved or even solvable. The finding of high item difficulty in particular poses some problems: items are considered difficult if they do not yield a reasonable number of affirmative answers (Moosbrugger 2008; Schermelleh-Engel & Werner 2008). Thus, an item that is not often selected as true will be eliminated from analysis although it might contain valuable information, e.g., that the statement is considered false by the majority of participating subjects. Further, in the case of subjects awakened from sleep, they may not affirm an item although it is true, simply because they are not yet able to comprehend its content (sleep inertia). For example, the item “I wasn’t in the dream, I had no self” was not often selected as true. Was this because in most cases, dreamers felt they did have a self or was it because they didn’t understand what was asked of them? I hope that this example highlights some of the problems that arise when we try to subject philosophical theory to experimental testing. Perhaps a different design, opting for a specific comparison of questions addressing epistemology vs. phenomenology during a steady state of wakefulness (such as mind-wandering or meditation) might generate more concrete answers, avoiding sleep inertia effects should they exist. We look forward to such results.

3 Insight as a state of consciousness?

According to Kühle, our results suggest that insight may be considered a state. Moreover, she claims that the LuCiD scale does not allow for the identification of different lucidity

levels. These assumptions are not to be deduced from our research but must stem from a misconception or misunderstanding of the factorial structure of the LuCiD scale. Concerning this matter, we reported that dream consciousness can be described by eight factors, six of which are capable of distinguishing between lucid and non-lucid dreams: insight, control, dissociation, positive emotion, negative emotion, and memory. A person can have a range of scores in each factor, for example in insight, such that scores are graded and allow for varying degrees of lucidity. Furthermore, the factors identified are correlated, i.e., not independent (see [Voss et al. 2013](#)), which means that one factor alone may not be sufficient to define a lucid dream. Our results also suggest that a dream might be considered lucid even with low scores of insight! So the assumption that the state of lucid dreaming is equivalent to the proposed state of insight cannot be inferred from our data. Kühle's proposal reveals another problem, however, that we tried to address with our Space of Consciousness model (SoC), which is the definition of "state". What is the relationship between a state of arousal and a state of consciousness? In the case of insight, the recognition "I am dreaming" may be only a fleeting thought. But this thought is embedded in relatively enduring neurophysiological patterns such as regional changes in blood oxygen levels (see [Dresler et al. 2012](#)) and enhanced gamma activity in frontal regions ([Voss et al. 2009](#); [Voss et al. 2014](#)). Our suggestion to situate lucid dreaming within the SoC attempts to incorporate these observations. In my view, a state is comparable to background activity enabling or disabling certain transients such as thoughts or memories. It is courageous to consider a fleeting thought a state, and I think such definition would need more detailed specifications. Of course, one may ask whether a dream would be considered lucid even in the absence or perhaps following the thought "this is a dream". According to our model, this assumption would have to be affirmed. If the state of lucid dreaming is considered a neurophysiological state of sleep bor-

dering wakefulness, enabling the mind to produce a transient thought (insightful thought), this thought may or may not be repeated several times within the state of lucid dreaming. The important factor is, as Kühle proposes, capability. During the state of lucid dreaming, the mind is able to be insightful. It is not the other way around, such that the mind is able to enter a lucid dream during the thought of insight. The importance of insightful thought thus does not lie in its being a state but in it being measurable! We cannot expect a subject to provide a truthful answer to the question "were your frontal lobes producing gamma band activity?" We can, though, ask about the quality of their thoughts and sensations. Finally, if, in spite of my objections, we define insight as a state of consciousness, how would this state be defined in terms of arousal (see the SoC model), or in terms of other determinants such as, for example, judging, sensing, or moving? Supposed insight were defined as a point in the SoC. Where would it be located? Within mindwandering, meditation, lucid dreaming, focused attention—or all of these?

4 Conclusion

While Kühle's comments are greatly appreciated, they show how important dialogue between the different disciplines involved in studying consciousness really is. Neuroscience, psychology, and philosophy are all connected in their quest for a better understanding of the true nature of consciousness and its underlying physiology. They depend on each other to formulate predictions based on theory, and to test and reappraise these on the grounds of cause-and-effect relationships established through experimental testing. Experimental research rests upon certain assumptions that may not or may only fractionally apply to philosophy. The most important assumptions of dream science are to consider it true that there exists a real world (1), that REM sleep dreams exist (2), that healthy awake humans are able to make valid statements about knowing and feeling (3), and that restrictions to this ability (e.g., sleep inertia) can be reliably identified (4).

References

- Dresler, M., Wehrle, R., Spoormaker, V. I., Koch, S. P., Holsboer, F., Steiger, A., Obrig, H., Sämann, P. G. & Czisch, M. (2012). Neural correlates of dream lucidity obtained from contrasting lucid versus non-lucid REM sleep: A combined EEG/fMRI case study. *Sleep*, 35 (7), 1017-1020. [10.5665/sleep.1974](https://doi.org/10.5665/sleep.1974)
- Duncker, K. (1947). Phenomenology and epistemology of consciousness of objects. *Philosophy and Phenomenological Research*, 7 (4), 505-542.
- Moosbrugger, H. (2008). Item-Response-Theorie (IRT). *Testtheorie und Fragebogenkonstruktion* (pp. 215-259). Berlin, GER: Springer.
- Schermelleh-Engel, K. & Werner, D. P. C. (2008). Methoden der Reliabilitätsbestimmung. *Testtheorie und Fragebogenkonstruktion* (pp. pp.113-133). Berlin, GER: Springer.
- Voss, U., Holzmann, R., Tuin, I. & Hobson, J. A. (2009). Lucid dreaming: A state of consciousness with features of both waking and non-lucid dreaming. *Sleep*, 32 (9), 1191-1200.
- Voss, U., Schermelleh-Engel, K., Windt, J., Frenzel, C. & Hobson, J. A. (2013). Measuring consciousness in dreams: The lucidity and consciousness in dreams scale. *Consciousness and Cognition*, 22 (1), 8-21. [10.1016/j.concog.2012.11.001](https://doi.org/10.1016/j.concog.2012.11.001)
- Voss, U., Holzmann, R. ., Hobson, A., Paulus, W., Koppehele-Gossel, J., Klimke, A. & Nitsche, M. A. (2014). Induction of self awareness in dreams through frontal low current stimulation of gamma activity. *Nature Neuroscience*, 17 (6), 810-812. [10.1038/nn.3719](https://doi.org/10.1038/nn.3719)

Representationalisms, Subjective Character, and Self-Acquaintance

Kenneth Williford

In this study I argue for the following claims: First, it's best to think of subjective character as the self-acquaintance of each instance of consciousness—its acquaintance with itself. Second, this entails that all instances of consciousness have some intrinsic property in virtue of which they, and not other things, bear this acquaintance relation to themselves. And, third, this is still compatible with physicalism as long as we accept something like *in re* structural universals; consciousness is a real, multiply instantiable, natural universal or form, but it likely has a highly complex, articulated structure, and “lives” only in its instances. In order to make these cases, I give a characterization of subjective character that accounts for the intuition that phenomenal consciousness is relational in some sense (or involves a subject-object polarity), as well as the competing and Humean intuition that one of the supposed relata, the subject-relatum, is not phenomenologically accessible. By identifying the subject with the episode or stream of consciousness itself and maintaining that consciousness is immediately self-aware (“reflexively” aware), these competing intuitions can be reconciled. I also argue that it is a serious confusion to identify subjective character with one's individuality or particularity.

I argue that deeper reflection on the fact that consciousness has only incomplete self-knowledge will allow us to see that certain problems afflicting acquaintance theories, like the one I defend, are not the threats to certain forms of physicalism that they might seem to be. In particular, I briefly consider the Grain Problem and the apparent primitive simplicity of the acquaintance relation itself in this light.

Keywords

Acquaintance | Consciousness | Direct realism | First-order representationalism | For-me-ness | Harder problem | Heidelberg school | Higher-order representationalism | Individuality | Individuation | Intrinsic property | Mineness | Naturalize | Particularity | Phenomenal consciousness | Phenomenal intentionality | Physicalism | Qualitative character | Reflexive awareness | Reflexivity | Relational property | Representation | Representationalism | Same-order representationalism | Self-acquaintance | Self-knowledge | Self-representation | Sense of self | Sense-datum theory | Stream of consciousness | Structural universals | Subject | Subjective-character | The grain problem | Transparency intuition

1 Introduction

In this study, I argue for the following claims: First, it's best to think of subjective character as the self-acquaintance of each instance of consciousness—its acquaintance with itself.¹

¹ As will become clear shortly, contrary to ordinary ways of speaking, I do not hold that persons must be the “subject relata” of acquaintance relations. Rather, I hold that episodes of consciousness are, fundamentally, the subject relata.

Author

Kenneth Williford

williford@uta.edu

The University of Texas
Arlington, TX, U.S.A.

Commentator

Tobias Schlicht

tobias.schlicht@rub.de

Ruhr-Universität
Bochum, Germany

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

Second, this does indeed entail that all instances of consciousness have some internal relational property (or intrinsic property) in virtue of which they, and not other things, bear this acquaintance relation to themselves. And, third, this is still compatible with physicalism as long as we accept something like *in re* structural universals. There is always a price, but in this case

it's arguably no more than the price we pay to be scientific realists.²

To make these cases, I must consider some important preliminaries. I give a characterization of subjective character that accounts for the intuition that phenomenal consciousness is relational in some sense (or involves a subject-object polarity), as well as the competing Humean intuition that one of the supposed relata, the subject-relatum, is not phenomenologically accessible. If the latter is true, it is hard to explain how we could have immediate evidence (as opposed to some sort of inferential knowledge) of the existence of this relational structure—evidence we do seem to have. If we identify the subject with the episode or stream of consciousness itself (however we individuate or ontologize these)³ and maintain that consciousness is immediately self-aware (“reflexively” aware⁴), then the intuition of relationality and the Humean intuition of the missing subject can be reconciled.

I also argue that it is a serious confusion to identify subjective character with one's individuality or particularity. This will be considered first from a phenomenological point of view, in relation to our tendency to describe subjective character in terms of ownership or “mineness”, and then from an ontological point of view, in relation to the metaphysical individuation conditions of distinct streams of consciousness.

Further, I argue that deeper reflection on the fact that consciousness has only incomplete self-knowledge will allow us to see that certain problems afflicting acquaintance theories, like the one I defend, are not the threats to certain forms of physicalism that they might seem to

be. In particular, I briefly consider the Grain Problem⁵ and the apparent primitive simplicity of the acquaintance relation itself in this light.

Preliminary to all this, we must first briefly consider the inadequacies of representationalism, and at least adumbrate some of the motivations for the recently renewed interest in the idea of acquaintance (see e.g., Chalmers 2003; Tye 2011, pp. 96–102; Gertler 2011, pp. 87–128, 2012; Balog 2012; Howell 2013, chs. 3 & 4; Goff forthcoming). I argue that, indeed, we need to lose our fear of moving beyond reductive naturalistic representationalisms, especially in regard to subjective character. My conclusions, and in many cases arguments, are not entirely new, but I attempt to cast the material in a new light, in a spirit of synthesis.

The dialectical structure of this study is somewhat circuitous. In section 2, I argue that the most plausible representationalist theory of consciousness is a self-representationalist one (or “Same-Order” representationalism) because it captures subjective character, which I view as essential to consciousness, with the smallest theoretical cost. However, I argue, all forms of representationalism about consciousness are ultimately implausible. This leads to a focused discussion of the notion of subjective character in section 3, the notion that motivates higher-order and same-order representationalisms. In that section, I argue that subjective character should be identified with the self-manifestation or self-appearance of consciousness. Consciousness, the claim goes, appears to *itself* no matter what else appears to it. This in turn allows us to make sense of the competing relationality and Humean “no-self” intuitions mentioned above. Combining these elements from sections 2 and 3, I argue in section 4 that we should understand self-manifestation in terms of self-acquaintance rather than self-representation. In section 5, I clear up what I regard to be the not uncommon confusion of subjective character with individuation. And in section 6, I argue

² This is not to imply that scientific realism entails physicalism, of course.

³ This is a difficult issue I will not enter into. See e.g., Dainton (2000, 2008); Strawson (2009).

⁴ I will occasionally use the terms “reflexivity” and “reflexive awareness” to denote just this characteristic of consciousness (i.e., that of its always being aware of itself). It is not to be confused with “reflection” in the sense of introspection. It is more like the logical usage of “reflexive” (as in “reflexive relation”). The acquaintance relation is reflexive on the domain of conscious states, according to the view accepted here (as well as being anti-symmetric). But not everything that stands in this relation is self-acquainted—episodes of consciousness are, but they are also acquainted with sensory qualities, and these latter are not acquainted with anything.

⁵ The Grain Problem, customarily attributed to Wilfred Sellars, is a problem for any identity theory according to which sensory qualities are really brain properties of some sort. Roughly put, the problem is that brain properties are complex and structured while sensory qualities seem, on the face of it, ultimately simple and unstructured. For good discussions with references to Sellars see Clark (1989) and Lockwood (1993).

that though the view espoused here implies that being conscious is a matter of having certain intrinsic properties, this is compatible with a certain type of physicalistically acceptable hylo-morphism—the view that complex kinds of physical objects, properties, or processes involve the concrete instantiation of real structures and cannot be properly understood in abstraction from such a “marriage” of form and matter.

2 Representationalisms: From first-order to same-order

In the theory of consciousness, the term “representationalism” has, aptly but somewhat confusingly from a historical point of view, come to designate any view according to which being phenomenally conscious is equivalent to representing the right sort of things in the right sort of way. There is, of course, much internecine disagreement over these things and ways, but the main idea is simple and attractive enough. If we could understand consciousness in terms of representation and representation in terms of some naturalistically acceptable relations, then we could “naturalize” consciousness. I’ll call representationalisms that are coupled with naturalistic theories of content *reductive representationalisms*.

Representationalisms are typically divided up into various “orders.” These orders have, in a way, to do with the kind of content (or object) a conscious representational state supposedly must have. For First-Order (F) representationalisms the relevant states are, fundamentally, just directed at worldly objects and properties (typically the sensible properties of tables, chairs, etc., see, e.g., Tye 1995, 2000; Dretske 1995). For Higher-Order (H) representationalisms, the states must be directed at mental states of “lower-order”—possibly but not necessarily first-order (see e.g., Rosenthal 2005; Lycan 1996). For Same-Order (S) representationalisms, the representational state must be directed at itself (or, perhaps, some part of itself, or a whole of which it is a part, or another part of the whole of which it is a part).⁶ I also add Priv-

ileged-Object (P) representationalisms as a distinct category. For these, the state must be directed at some special type of entity—a model of the organism as a representational or embodied homeostatic system, a “proto-self” or, less naturalistically, perhaps an enduring substantial ego entity.⁷

There is, however, no obvious reason why there could not be unconscious representations with any of these contents. And, generally, it seems implausible that something could be conscious in virtue of representing a certain type of object—this is Alvin Goldman’s so-called “Problem of the Rock” (thinking about or seeing rocks does not make them conscious, so why should it make anything else conscious?), which seems to apply to H, S, and P theories—but see below (see e.g., Goldman 1993; Gennaro 2005; Lycan ms).

For F theories, since it is admitted there can be conscious and unconscious states with the same sort of content, another distinguisher between conscious and unconscious mental states will have to be found. For F theorists, this has typically been a functional constraint placed on the representations (e.g., poise, feeding into the mind-reading system, becoming available to the global workspace, see, e.g., Tye 2000 and relatedly Baars 1997; Dehaene & Naccache 2001), sometimes coupled with the necessary condition that the properties represented must be represented in a “non-conceptual” way (whatever that is taken to amount to).⁸ For the H theorists, it has been a somewhat different story.

H theorists are generally motivated by a phenomenological inadequacy they see in F the-

theorist. S theory is also often called self-representationalism.

7 For naturalistic versions, see e.g., Damasio (1999 and 2010), Metzinger (2004), and Sebastian (forthcoming). I am sure that Damasio, Metzinger, and Sebastian would reject this label, but the point of it is that all these theories identify subjective consciousness, in one way or another, with the representation of a “self,” understood in a naturalistically acceptable sense. See e.g., Metzinger (2004), p. 302: “In short, a self-model is a model of the very representational system that is currently activating it within itself” (emphasis original); and Damasio (2010), p. 180: “[T]he brain constructs consciousness by generating a self process within an awake mind. The essence of the self is a focusing of the mind on the material organism that it inhabits.” It should be noted that Metzinger allows that there could be conscious experience that does not involve subjective character (see Metzinger 2004, pp. 559–560). Thus my categorization here applies at most only to his theory of subjective consciousness. Since, for me (as for Damasio), all consciousness necessarily has subjective character, this difference in detail will not loom large in what follows.

8 See the excellent discussion of the “non-conceptual content” literature in Hopp (2011).

6 See e.g., Gennaro (2012); Kriegel (2006, 2009); Weisberg (2008, 2014). Williford (2006) can be taken to express a pure S view—the conscious mental state has itself for its own object, not some portion of itself. We can also classify Carruthers as an S theorist; see Carruthers (2000, 2005). Gennaro would not describe himself as an S

ory. F theorists generally stress the so-called “Transparency Intuition”—the idea, roughly put, that first-order consciousness reveals only properties and objects in the world and nothing directly about consciousness itself, the perceiving mind, the subject, or the vehicles of representation (see e.g., Harman 1990; Tye 2000; Byrne 2001). H theorists, on the other hand, are with varying degrees of explicitness motivated by the equally powerful intuition that consciousness involves some sort of “for-me-ness” or “to-me-ness,” often termed “subjective character” (see e.g., Rosenthal 1986, p. 345 and Gennaro 2006. See also Levine 2001, pp. 104–111). This gets encoded in the H mantra that the conscious states are just those that one is “Aware of Being In”, those that one is aware that one is oneself in (see e.g., the “Introduction” to Kriegel & Williford 2006). The thought is that F theory simply does not capture that intuition. F theorists and their fellow travelers would consider such “essentially indexical” contents or the “sense of self” to be more advanced cognitive products or artifacts of social cognition, certainly not in the very ground floor of consciousness (see e.g., Edelman & Tononi 2000, pp. 103–104 and Macphail 1998, pp. 2–5).

There is here an important bifurcation in intuitions about consciousness. Some significant percentage of us thinks that subjective character (however we ultimately understand it) is essential to consciousness, is in the ground floor. And some significant percentage of us thinks that it is not; that somehow qualitative character (perhaps understood as having the right sort of representational content) is essential but that subjective character is derived, secondary, or tertiary. This bifurcation shows up in neuroscientific and psychological thinking on consciousness as well.⁹ We will briefly return to the significance of this bifurcation point in the next section.

The H theorist has a few options about the exact content of the H representation, the higher-order thought (or perception [or global state]).¹⁰ There are serious and well-known

problems here. If the represented lower-order state (L state) of, say, visual perceptual awareness were different from the representing H state in terms of relevant content (e.g., if the one represents a phenomenally green ball and the other a phenomenally red one), what would we consciously see? “Red. No, green. No, red...” This is the Problem of the Division of Phenomenal Labor, or mirepresentation problem, as I will sometimes call it, and is related to deep and probably insoluble problems about the epistemology of introspection that are pertinent to such models (of both H and S varieties).¹¹ If the L state simply did not exist, would your conscious experience in that case be a sort of Meinongian hallucination? This is the Problem of Targetless H States.¹²

To take up the latter problem just a bit, if one takes literally much of the talk one finds in the literature on H theory, the H thought is supposed to *make* the L state conscious. Being conscious is a kind of extrinsic (external relational) property of the L state, a property it has in virtue of its being represented by the H state. Thus, if there is such an H state, it does confer at least a relational property (the property of “being made conscious by the H state”) on the L state. In the cases in which the L state does not exist but the H state directed at it does, some non-existent object, the L state, is made conscious by an H state. Thus the L state would literally have a relational property; it would stand in a relation, even though it does not exist. This literal interpretation of the view entails some form of Meinongianism (at least about non-existent L states) and that you can seem to yourself to be conscious when you are not. Thus, presumably, it should not be taken so literally.¹³

¹¹ See e.g., Neander (1998); Horgan & Kriegel (2007); Weisberg (2008); Tye (2011, pp. 4–8). See Kidd (ms) for an excellent discussion of these epistemological issues in the (not interestingly different) case of S theory.

¹² See Mandik (2009 and forthcoming) on the “Unicorn problem” and Block (2011). See Rosenthal (2011, 2012); Weisberg (2011a, 2011b); Kiefer (2012); Wilberg (2010), and Berger (2013) for discussions of various strategies for dealing with Higher-Order-Thoughts (HOTs) without Lower-Order-Thoughts (LOTs).

¹³ What I am calling the “non-literal” interpretation is, in effect, the position in Berger (2013). And in Rosenthal (2011, p. 436) he in effect claims that the non-literal position (as I am calling it) has always been his view. See Mandik (forthcoming) on this.

⁹ For example, Tononi & Koch (2008, pp. 240–241) do not seem to think that the “sense of self” is essential (though Tononi (2014) may have recently changed his view); Damasio (1999, 2010) is in the opposing camp; see also Northoff (2013).

¹⁰ I’ll not go into the Higher-Order Thought vs. Higher-Order Perception debate. See e.g., Gennaro (2012).

The non-literal interpretation, however, is inimical to one of the reductive pretensions of the H strategy. It's not inimical to reductive representationalism as such. But it does draw in to question the idea that a reductive theory must construe the property of being conscious as an external-relational property of otherwise unconscious mental states (see [Rosenthal 1997](#)). Thus, it could only be in virtue of the specific content or structure of the H state itself that there is consciousness. One would then be putting forth the presumably phenomenologically motivated *a posteriori* identity hypothesis that the conscious representational states are just the ones with that content. There may be differences over the specific content (e.g., Is it about some of my other mental states, or is it just about the non-mental objects and properties of the world?) and differences over other criteria (e.g., poise); but otherwise, on the non-literal interpretation, H theory is structurally just like F theory. We can of course wed either of these to a reductive theory of representation, but this will only make "being conscious" into an external-relational property to the extent that the theory of representation adopted makes all representation an external-relational matter.

If one is still conscious when the L state does not exist, then the H state would seem to be doing all the work. And that's what we should focus our explanatory efforts on. What could be special about it? Again, putting aside other types of external relations (e.g., being available to the global workspace), it must have a special sort of content. But it is not in virtue of *being represented* that a state could be conscious. Rather, on this non-literal interpretation, it is in virtue of *being a representation of X* (where *X* is a special object of some sort, e.g., oneself being in a state) or *that p* (where *p* is a proposition with a special content) that the state is conscious; and we can, as with any other sort of contentful state, try to figure out how different naturalistic theories of representation would construe states with that content.

Whatever theory of content we adopt, we'll want to know what salient or interesting properties, from an explanatory point of view, such representations have. What is it about you

that you can represent yourself as being in a state or that a conscious state of yours is occurring now? Find that out, the promise goes, and we will understand consciousness. But, I would argue, none of the theories of representation we have to go on tell us anything very significant about such states. The beaver's tail splash, says Millikan, to take one sort of example, can represent the very time at which it occurs (among other things; [Millikan 1995](#), p. 98). This does not make it conscious. This particular example applies directly to Same-Order theories, but surely the beaver's tail splash *could have* represented a previous tail splash and its content or its simultaneous front paw splash, etc., but that would not in itself make anything conscious either, right?

Naturalistic theories of representation will not themselves tell us anything that interestingly distinguishes H states (or S states) from F states (or P states for that matter). In every case (F, H, S, P), it is just a matter of some physical representational vehicles standing in some set of external (or externally mediated) relations to other physical objects (and sometimes to themselves). From this point of view, we see nothing that interestingly distinguishes the theories.

Moved by these problems, H theorists might try to go the "essential indexical" route (cf. [Weisberg 2012](#)). After all, on Rosenthal's original formulation, the conscious states are those one is aware of *oneself* as being in. But here they are faced with a difficult choice. If they presuppose a teleosemantic theory, then they have to face the fact that on this theory there are no literally *essential* indexicals (see e.g., [Millikan 1990](#)). Change the relevant history and other external relations and you change the content—now an indexical, now a proper name, now a substance term, etc. If they abandon teleosemantics, they could go back down some Fregean rabbit hole.¹⁴ That way lies murk or perhaps triviality (see [Cappelen & Dever 2013](#)). But it seems inadequate just to postulate that

¹⁴ I assume here but will not argue that teleosemantics is the most plausible naturalistic theory of content. There may be other naturalistic options that allow one to make good sense of the notion of essential indexicality in a way that could help H theory here, but I doubt it.

the H state contains a definite description that happens to pick oneself out. Thus the H theorist might be led to consider what is in effect a P theory. One then tries to find a suitable entity to play the role of the privileged object (a privileged signified, if you will): the proto-self, the self-model, or what have you.¹⁵

It is hard to see how any of these possible objects would somehow help us to make sense of subjective character. And it is hard to see how representing some special object could be that in virtue of which something is conscious. If “essentially indexical” content is either explicable in terms of something more basic (as seems to be the case to me), or impossible (as on teleosemantics), or metaphysically fraught in an ultimately un-illuminating way, then it seems like the best bet is to adopt a version of S theory.

For one thing, we can reduce the metaphysical load that threatens to plague the notion of essential indexicality and solve the non-existence problem at once.¹⁶ All we need are token mental states representing themselves. As a corollary, we can give a deflationary account of “essentially indexical” content in token-reflexive terms¹⁷ that is potentially compatible with teleosemantics (or whatever non-Fregean account one prefers) and find some other way to capture the grain of truth reflected in the opacity arguments presented by Castañeda, Lewis, and Perry.¹⁸ In my view, anyone committed to the intuition motivating H theory should become an S theorist, if for no other reason than because of the non-existent L state problem. The other possible solutions (e.g., Gennaro’s “WIV”) introduce a kind of theoretical inelegance that renders them less plausible.

H theories are better than F theories, given my intuitions anyway, because they en-

code the essentiality of subjective character to consciousness. If that intuition is good then, of the two classes, H theories are the better ones. But H theories face the non-existent L state problem. To solve it, they must either embrace murk or metaphysical baggage (if they go in the direction of some P theories), or embrace the postulation of certain epicycles, or go some other order. S, in my view, is evidently the best option for the representationalist.

S theory avoids *ad hoc* moves, better reflects the clarified phenomenological intuitions that are the real motivation, can ground a theory of indexicals, and does not commit one to an enduring self-entity of any sort; nor does it seem to attempt to get subjective character out of something’s representation of something else that is structurally similar to itself, as this last move runs afoul of the Fichte-Shoemaker Regress.¹⁹ S theory evidently does not fall prey to the non-existent L state problem, even if it does not avoid the misrepresentation problem. In the end, however, it is itself nothing more than a type of P theory. The Privileged Object is just the token mental state (or episode) itself. Clearly, there is no self-evident reason why something’s representing itself should make it conscious, even if it is in fact true that all conscious episodes do represent themselves.

We surely cannot seriously imagine that consciousness emanates from a special object it needs to look at, even if that object is just the current experiential time-slice itself. Further, something’s representation of itself, naturalistically understood, is no more theoretically inter-

¹⁹ See Henrich (1982); Frank (2002, 2007); Shoemaker (1968). The issue, which is part of the “essential indexical” problematic, is, when put into a “self-model theory” context (which is not to be identified with Metzinger’s views), just that modeling something structurally isomorphic to oneself is not sufficient for knowledge that one is modeling oneself, as opposed to having behavioral control through such an interface (I could be controlling my doppelgänger unwittingly and just as effectively). One would need to know that the thing modeled is oneself (and not something else that happens to be isomorphic to it, like one’s counterpart in a close possible world). One cannot, on pain of regress, derive such knowledge from a set of descriptions of oneself without already knowing that at least one of the descriptions does indeed apply to oneself. So one must have some direct self-knowledge, such as knowledge by acquaintance that one is the relevant so-and-so. An S theory wedded to a teleosemantic theory of representation and externalist theory of justification has the advantage of being able to accommodate direct reference and non-inferential knowledge of oneself, though one will regard this as a mere simulacrum of the phenomenology.

¹⁵ See Sebastian (forthcoming) for a Damasio-inspired turn toward a P theory (at least, that was my interpretation of it).

¹⁶ We can’t eliminate the misrepresentation problem, however. But we bracket that for now. See Kidd (ms) and Weisberg (2008).

¹⁷ A la Higginbotham (2003 and 2010) and before that (implicitly) Smullyan (1984); see Cappelen & Dever (2013, pp. 160-161). The hyperset model in Williford (2006) is the skeleton of such a theory. See also Kapitan (2006).

¹⁸ See Cappelen & Dever (2013, ch. 10). They attempt to capture this grain by appealing to relatively un-puzzling epistemic limitations. I believe they are on the right track, even if I would characterize the specific limitations in question a bit differently (see the discussion below on our ignorance of what fundamentally individuates us).

esting (or even surprising) than its representation of the world or of one's other thoughts and perceptions. Thus, that does not, *a priori*, appear to be the sort of thing that would be more likely to be equivalent to consciousness than something's representation of something else. Perhaps adding functional constraints would help here but no more than it might help H or F theory.

Even if it is true that all conscious states are self-representational, it is, of course, far from clear how that fact should help us *explain* consciousness. The same can be said for H theory and other P theories. Rather, in all these accounts, we are merely trying to isolate what we take the unique content of consciousness to be and then to apply our theory of representation to states with such content. Absent some strong phenomenological intuitions to the contrary, the conscious mental states, it seems, might well have been all and only those states in which dogs are represented. In the end, though, all "normal" physicalists (i.e., those who reject Russellian Monism, Panpsychism, and Pan-proto-psychism) are reduced to *some* such strategy. All "normal" physicalists, representationalist or not, will identify consciousness with something that is not *a priori* known to be equivalent to it. We return briefly to this familiar problematic at the end.

But, perhaps most alarmingly, reductive S theories (and H theories, and everything in between) are either subject to a version of the old Swampman objection or otherwise untenable.²⁰ Since the conscious states are, on the theory, just special representational states, they are subject to the constraints of the underlying theory of representation (in this case, teleosemantics). If they don't have the right history, then they don't have the right content. And if they don't have the right content, they are not conscious. Surely there is something simply absurd about the idea that one might or might not be conscious depending on how one's atoms happened to get into the current arrangement.

²⁰ See e.g., Tye (2000, ch. 6). I will not be able to go into the back and forth over Swampman. Suffice it to say that despite hearing many attempted rejoinders over the years, I still find the objection to be a *reductio* of representationalist theories of consciousness wedded to historico-externalist theories of content.

It is not that one cannot concoct a response to the objection; it is, rather, just the very fact that the view invites such objections in the first place. It demands a rather serious and ugly epicycle; and that counts strongly against it. But if we reject teleosemantics and adopt an internalist theory so as to escape from Swampman, we face equally difficult problems that we cannot, unfortunately, go into here.²¹

The view then is that H theories are better than F theories on phenomenological grounds and that S theories are better than H theories on dialectical *and* phenomenological grounds. But all versions wedded to naturalistic historico-externalist theories of representation are shipwrecked on the Swampman problem, and internalist versions face other equally difficult problems. What then shall we do?

We might consider trying out a non-reductive representationalist version of S theory. This is a possibility we will return to in section 4. But first we need to reflect a bit on what H, S, and P theories are trying to capture in the first place. What is the phenomenological datum designated by this phrase "subjective character," and why is it that F (and related) theorists don't see it as essential to consciousness, while H, S, and some other theorists do?

3 Subjective character

Subjective character is often described as a certain "for-me-ness," "mineness," or even "me-ishness" that is phenomenologically manifest and, presumably, always accompanying, even if in a muted or background form, any consciousness whatsoever (see e.g., Zahavi 2005; Levine 2001; Kriegel 2009 and Block 1995). F and related theorists point out that it also seems that one can become so absorbed in one's actions, at one extreme, and perhaps so dulled at the other that one loses all sense of oneself (see e.g., Tononi & Koch 2008, pp. 240–241). Moreover, they might argue that it does not seem reasonable to suppose that worms and bees have a

²¹ See e.g., Carruthers (2000, 2005) and Gennaro (2012, pp. 45–49). Briefly, the sort of functional role semantics Carruthers embraces derives actual, occurrent content from dispositions, and it is actually subject to variations on the Swampman theme.

sense of self at all, and yet they may be conscious. A common reply from the defenders of subjective character to the first claim is that we are not talking about focusing on oneself or one's current mental state as an object of attention or concern, and that, if they tried harder, F theorists would realize that even in the most dulled or, at the other pole, absorbed state, they are still aware at some level of themselves (or the very experiential state they are in). To the second objection, the typical reply is that the sort of subjective character we are envisaging does not require the sort of conceptual sophistication or reflective capabilities that would make it impossible for dogs (or even bees and worms) to count as conscious beings (see e.g., [Gennaro 2012](#), chs. 7 & 8). Of course, the replies can be replied to, and so on. And we won't enter into these debates here. Suffice it to say that, unsurprisingly, those who think that subjective character is essential to consciousness have ways of answering objections, just as do those who deny its essentiality. As commonly happens, the answers drive us back to questions that are themselves at least as hard to settle as the ones we began with. Moreover, appeals to the neuroscientific and psychological literature in the attempt to decide these issues sometimes get what plausibility they have from interpretations of the experiments and results that are as questionable as the claims they are supposed to support.

My view here is that one should follow the modeling path inspired by one's "phenomenological muse" and give up fighting phenomenological intuition wars. If you find subjective character to be essential, develop models of consciousness that encode that, and see where they lead. If you don't find it essential but find other things to be more important (multimodal information integration or availability in the global workspace or whatever), model those. And let's not forget that we might all be working on different parts of the same elephant, so perhaps we will be able to combine models fruitfully one day. Eventually we may have ways of more or less decisively testing the different models.²²

²² See [Kriegel \(2007\)](#) for an excellent discussion of phenomenological impasses. Thanks to Jennifer Windt for reminding me of this lucid article.

Different intuitions about what is essential to a phenomenon drive different models of the phenomenon. As long as enough people (and don't ask for a number) share one's phenomenological intuitions, one's project won't be, we hope, insane or unmotivated. In regard to the present bifurcation point, many otherwise sane, rigorous, and careful thinkers in many widely distributed traditions and disciplines have had some version of the intuition that consciousness, somehow, involves a sense of self or sense of itself.²³

Now, how should we characterize subjective character at the phenomenological level? It does not add much to say that it is a "sense of self." What sort of a sense of self are we talking about? To say that it is "mineness" or "for-meness" makes it seem as though we are talking about the *ownership* of experiences. But this is probably just a certain analogy based on the ownership of property. Yes, for all that matters here, it may well be the case that, always, if I am in a position to know, without having to observe any behavior, that *there is* a pain in the room, then I am in a position to know that it is my *own* pain in the room. But it does not do much good to say that "me-ishness" or "mine-ness" adheres to my experiences like a property or haecceity. It is not as if I just see that my experiences have Willifordhood instead of Zahavihood or Gallagherhood, and thereby know whose are whose—like distinguishing two otherwise qualitatively identical coats by different name tags on the inner pockets.

Note that looking for a special property of the experience is not that different from seeking out its relation to a special object (its owner or The Self) that one may be directly acquainted with. In both cases we are looking for a special something that individuates the experience. There is no interesting difference here between a special unique property that only my experiences have and a special unique self-object to which they all relate.

²³ For just a few examples of the historical pedigree here, see [Caston \(2002\)](#) on Aristotle, [Williams \(2000\)](#) and [Coseru \(2012, ch. 8\)](#) on the Indian and Buddhist debate, [Thiel \(2011\)](#) on the early modern problematic, [Frank \(2004\)](#) on the German Idealist and Romantic discussion, and [Zahavi \(1999 and 2006\)](#) on the Phenomenological movement.

Subjective character should probably not be thought of as a matter of a constant relation to a self-object or as a special property of mine-ness or me-ishness that all experiences come with, all the more is this so if it is possible to misattribute ownership to certain sensations.²⁴ The first-personal dimension (Zahavi), the sense of self in the act of knowing (Damasio), for-me-ness, me-ishness (Block), ipseity, *être-pour-soi* (Sartre), *Selbstvertrautheit*, and so on—these are all suggestive names for the phenomenon in question. But we'd like to know if there is not an at least somewhat less ambiguous way of characterizing it.

One name for it that I do rather like depends on a grammatical analogy that can be fleshed out a bit more. Every experience, we may say, involves the appearance *of* something *to* something (or someone). The former can be called the *genitive of manifestation* (appearance-of), the latter the *dative of manifestation* (appearance-to).²⁵ The genitive of manifestation corresponds to the intentionality of consciousness—its directedness at objects; the dative of manifestation corresponds to subjective character. The identification of subjective character and the dative of manifestation may not at first be so obvious.

The primary intuition here is that there is no such thing as the mere non-relational phenomenal appearance of an object or quality. Objects and qualities don't just phenomenally manifest—full stop. Rather, anything that phenomenally appears, appears to someone or something (cf. Strawson 2011, pp. 41–46). If this were false, phenomenal consciousness would be more like a monadic property of its objects than like a relation between a subject and an object of some sort (see Butchvarov 1979, p. 250). The idea that consciousness could be phenomenally manifest but manifest *to* no one is

either incoherent or, at best, strains credulity. Yet this seems to be exactly what F and related theorists are committed to—aches and pains that can appear (be phenomenally conscious) but appear to no one.

If we accept that there is a dative of manifestation, that objects and qualities appear *to* someone or something, we are closer to but not quite up to subjective character just yet. Subjective character, recall, is supposed to be something phenomenologically detectable. And one might raise the following sort of worry. Suppose phenomenally manifest objects and properties are manifest *to* something or someone. It does not follow from this alone that that *to which* they are manifest is itself manifest or even manifestable. Nor does it follow that the *fact that* they are manifest to something is manifest or even manifestable. In other words, there could indeed be a dative of manifestation and yet no direct phenomenological evidence of this at all. In fact, Hume's famous failure to find his own self and Moore's similar but more tentative musings on this issue can be taken as expressions of the intuition that we do not find a distinct subject relatum in experience.²⁶ And surely it is true that we do not find a little ubiquitous homunculus—the constant and ever-present thing Hume might have been seeking, like the little face at the bottom of old first-person video games like Quake—to which all our experiences relate—nor do we find a self-haecceity forever re-instantiated by our conscious episodes.

There is, however, this strong intuition that phenomenal consciousness is relational, that it involves a subject-object polarity. And the strong intuition that we do not find any entity or special criterial property that could be a self-entity, me-haecceity, me-ish quale, or subject-relatum is in some apparent tension with this intuition of relationality. Moreover, a *hidden* subject-relatum would not account for the phenomenology of subjective character, evidently. There is a real question here. How is it that consciousness seems to have a subject-object relational structure, and yet we do not

²⁴ See e.g., Lane & Liang (2011). (Thanks to an anonymous reviewer for pointing this nice article out to me.) If, as I shall argue, subjective character is not fundamentally a representational matter at all, the issue of *representational* immunity to error through misidentification is orthogonal. To the extent that the attribution of ownership is a representational matter, it may or may not be possible to misattribute ownership, as far as the view defended here is concerned.

²⁵ The terminology apparently derives from Prufer (1975) and is very common in phenomenological quarters. See e.g., Zahavi (1999); Crowell (2011, p. 16).

²⁶ See Moore (1910); Butchvarov (1979, p. 250, 1998, p. 55), and Williford (2004). On Hume in this regard, see Strawson (2011).

seem to be able to find the subject-relatum, one of the relata of the relation? Isn't it the case that if something non-inferentially seems relational, then we are non-inferentially aware of its (at least) apparent relata? Speaking naïvely and barring certain irrelevant counterexamples, if I see that the cup is on the table, don't I see the cup and see the table too? In the case of the subject-object polarity, do we imagine or project this relation? Is it a product of reflection and memory?

It seems to me that the F theorist should say that it is somehow a product of higher cognition that is projected onto normal adult human conscious experience. But if one is really committed to the intuition that subjective character is an essential and hence ubiquitous feature of conscious experience, then one will simply have to abandon self-relatum and self-haecceity accounts as characterizations of the phenomenology (and as explanatory models, for that matter). What we need is an account of how it is that consciousness manifestly and non-inferentially appears to have a relational structure even though one of the relata is, in a certain sense, invisible.

Here the view that consciousness is self-manifesting can save the day. An episode or perhaps stream of consciousness, on this view, appears to itself at every moment while other things appear to it as well. This will require more unpacking, but at present we just want to clarify the putative phenomenological content of the claim as best we can. We leave the notion of *appearing* or of *phenomenally manifesting* undefined. Or, if you prefer, we define it ostensibly by inner ostension and hope that our interlocutors know what we are talking about and have similar conscious minds (cf. Fales 1996, pp. 147–148).

Let's say that phenomenal manifestation is just the appearance to/in consciousness of something. Let's leave it open what that something is (qualities, facts, objects). We all can know what phenomenal manifestation is, in this purely phenomenological sense, if we are conscious and capable of normal reflection, attention, memory, and conceptual cognition. If we have tasted coffee, then the taste of coffee has

been phenomenally manifest to us. If we haven't, then it has not. And think of this generically—it's what experiencing the taste of coffee has in common with seeing the blue sky and with feeling one's own existence.²⁷ Now, the claim is that an episode of consciousness is phenomenally manifest to itself whenever anything else is phenomenally manifest to/in that episode. Whenever anything else appears to consciousness, that act or episode or stream of consciousness appears to itself as well. And it is important to remember that this does not mean that one is reflecting on one's experience or that one has any propositional attitude towards that experience or that one is paying any attention to that experience as such.

Now, let us suppose that this is the case. Can we recover a notion of subjective character from this in a way that accounts for both the Humean intuition that the subject-relatum is, in some sense, invisible and that, nevertheless, consciousness has a subject-object relational structure that is phenomenally manifest and non-inferentially knowable? Yes, we can, and at a relatively low price.

The subject-relatum, on the current proposal, is just the episode of consciousness itself. The episode appears to/in the episode. Other things (qualities, objects, etc.) appear in/to the episode as well. The episode is a unified whole, the differentiated qualities and objects appearing in/to it are like its parts (stressing "like"—it's an analogy).²⁸ We do not find episodes that do not have parts (except perhaps in some very special circumstances), but it is foolhardy to look for some special entity or haecceity that is separable from all the other parts or like a part among the parts. There is no such thing. And that, arguably, is the sort of thing Hume was failing to find. No such subject is given, hence we don't find it. Nonetheless, the true subject-relatum, the episode of consciousness itself, is not invisible. It is manifest.

²⁷ Cf. Moore (1910, p. 57). (This paper of Moore's is not as well known as his "Refutation of Idealism," but it deserves to be.)

²⁸ I will not attempt to offer an account of the (synchronic or diachronic) unity of consciousness in this paper (again, see e.g., Dainton 2000) or of mereological principles governing "parts" of episodes of consciousness and episodes as "wholes." It is enough for my purposes that one recognize that conscious episodes are internally variegated unities of some sort.

The main price to pay here is that we must try to wrap our heads around the idea that an episode of consciousness could be the phenomenological *subject* of consciousness. I say, and say truly, that such and such appears to *me* or that *I* see, feel, hear, or am conscious of such and such. If I am a subject of consciousness and all subjects of consciousness are just so many episodes, then am I just an episode of consciousness?! I've seen the incredulous stares with my own eyes and have been told that the sentence expressing the view that the subject of consciousness is the episode of consciousness has the same status as sentences like, "Pink dreams sleep furiously."

Indeed, this claim seems wildly counterintuitive at first. But once we realize that there is a certain temporal element connoted in our usage of "I," then this can be ameliorated. "I" normally refers not just to the present experience but to a whole history of connected experiences and much else besides. So it would be a mistake to infer from "I've seen the incredulous stares" the claim that "Incredulous stares were seen in/by this current episode of consciousness." Instead, in the spirit of Four Dimensionalism, one should translate thus: There was a past series of conscious episodes suitably connected to each other and to the present one; incredulous stares were seen by/in them for some time; and the episodes are being recalled in/by the present conscious episode, which bears the same relation (transitively conceived), or some suitable analogue thereof, in the case of broken streams, to that sequence of earlier episodes.

Note, however, that fundamentally the use of "I" is anchored in moment-by-moment, self-manifesting conscious experience. Imagine a person with severe anterograde amnesia and retrograde amnesia as well. Such a person might think, from moment to moment, "I am seeing this," "I am feeling that," but beyond a certain perhaps necessary amount of working memory, they may not carry any of that information into their future. We can imagine truly minimal subjects that have only the minimal amount of working memory required for consciousness, supposing that some amount is required. On the view proposed here, such a conscious being's

consciousness would still have subjective character. It would simply fail to be more or less automatically enriched by memory, projection, familiarity with one's body and dispositions, autobiographical idealizations and distortions, etc., that is, by the autobiographical representational grid through which our experience is normally spontaneously filtered. Perhaps such a person could not think "I" in the sense in which we normally think it. They may lack an "autobiographical self" and even "extended consciousness", as Damasio would put it (see [Damasio 1999](#) and [2010](#)). But their experience would be self-manifest and other things ("parts") would be manifest in/to that experience as well.

Still, isn't it a bit too odd to hold that the whole episode is conscious of its "parts"—however we end up construing these? Or that the "parts" are phenomenally manifest to/in the whole they belong to? Doesn't this still seem like a totally bizarre thing to say? We have to remind ourselves that there is no thing in consciousness, no ego entity, no homunculus that these qualities could be manifest *to*. We don't find any such thing; and no *hidden* thing could allow us to account for the phenomenology. However, we agreed (I hope) that consciousness has a relational, subject-object structure and that this structure is itself phenomenally manifest and not inferred.

Another way to put it is to say that there is a kind of contrast present in our experience all the time. Something is before me, and it is not me. Something is present to consciousness, but it is not that consciousness. Given our mereological analogy, this contrast is a bit *like* that between a whole and its proper parts. The whole is not a proper part. Yet, at a suitably generic level, it bears the same relation to itself that it bears to its constituents (everything is a part of itself too, though an improper part).

Assuming that this relational structure is not projected onto the experience in reflection, assuming that is, that this is a genuine "prepredicative" structure of experience, the contrast between the subject-pole and the object-pole is manifest, even if it normally remains unthematized or attended to as such. On the hypothesis that consciousness is always self-manifesting,

there is no problem here. The relevant contrast is *like* the contrast between the parts and their unified whole. The parts are manifest. The whole is manifest (self-manifest). So all the needed elements are present for their relations (of differentiation, unification, and inclusion) to be manifest.

Moreover, the idea that the difference between the parts and the whole is prepredicatively manifest is no more implausible than the idea that the difference between parts and other parts is prepredicatively manifest, something almost no one would deny. If I see a red patch on a black background, I have a differentiated, contrastive visual experience. The same goes for differences between the sensory modalities: we see and hear simultaneously, etc. If those sorts of contrasts can figure into the ground level of experience, why not the contrast between the unified self-manifesting whole and all its manifest “parts”—the totality of simultaneously manifesting qualities (however we understand them exactly) in all modalities (sensory and possibly cognitive, conative, and affective)?

Subjective character then, on this view, is just the self-manifesting character of an episode of consciousness. This view has the nice feature that it allows us to simultaneously account for the Humean-Buddhist “no-substantial-self” intuition and the intuition of relationality, with its attendant minimalist “sense of self”—as subject-pole.²⁹ It does this with less metaphysical cost than self-entity and self-haecceity theories, even supposing that those theories are not entirely phenomenologically implausible and explanatorily bankrupt. Let’s remember, however, that this is meant as a phenomenological claim fundamentally: consciousness is self-manifest just as the unified totality of sensory qualities (etc.) is manifest; and their contrast is manifest too, just as the contrast between such qualities (etc.) is manifest. This phenomenological claim has an ontological significance only if we accept that consciousness is indeed how it seems to be upon reflection. A claim that I accept in this case, but one need not accept it to appreciate

the phenomenological point and the virtues of this way of articulating it.

4 From self-representation to self-acquaintance

I gave up on reductive self-representationalism for quite general reasons, reasons affecting all representationalisms. As such, one might be tempted to suggest adopting some non-reductive form of S theory. For example, if one adopts the phenomenal intentionality³⁰ view, one might hold that whatever phenomenal representation *is*, consciousness represents itself in *that* way. It seems like this view might be just another way of describing the same phenomenological facts belabored in the previous section. If that is so, the phrases “phenomenal intentionality” and “acquaintance” are going to be basically synonymous, and the advocate of the former terminology can just translate. If we build nothing into the notion of representation other than the idea that something (an object, property, episode of consciousness, or whatever) is phenomenally manifest (to someone), then the views are indistinguishable at the phenomenological level and, maybe, the ontological level as well.

If this is not what is intended, however, then it is probably because the phenomenal intentionality theorist wants to mark an important distinction between intentionality (representation) and acquaintance. Perhaps they would prefer not to be committed to acquaintance if possible, and there are several reasons they might want to avoid such a commitment. But I will argue that in a certain sense, to be plausible at all, all forms of representationalism, reductive and non-reductive (including a phenomenal intentionality-based representationalism), ought to embrace a type of acquaintance relation.

Consider, for a moment, fictionalist representationalism about sensory qualities (projectionism about colors, for example). One could embrace a view according to which the sensory qualities are phenomenally manifest, though they in fact are never really instantiated by

²⁹ I defend this view also in Williford (2011a, 2011b) and in Williford et al. (2012); Dreyfus (2011) is an articulation and defense of a similar view from a Buddhist perspective.

³⁰ See e.g., Kriegel (2011) and the papers in Kriegel (ed.) (2013), as well as Kriegel’s excellent introduction to that volume.

anything. In such a case, one would not want to think of sensory phenomenal consciousness as a matter of bearing a real acquaintance relation to such qualities or quality instances. Instead one might prefer an adverbial construal of the situation that avoids any commitment to anything literally having (or perhaps even to there being) the properties phenomenally represented. On this view, one denies that there is a relation that supports existential quantification over these immediate objects (whatever they are), and one cannot conclude from the fact that one is phenomenally conscious of a red patch that there exists a red patch of which one is conscious.

Of course, this failure of existential quantification won't apply in the case of one sort of object, namely the conscious episode itself. But it will not be *because* it is an object of phenomenal intentionality that one can validly, existentially generalize from it; generally that fails, just as in other intentional (and intensional) contexts. Rather, it will be because it is the subject or *bearer* of phenomenal intentionality that one can validly generalize from it. In other words, we take episodes of consciousness to be individuals that have this pseudo-relational property. That is why we can quantify over them, and not because of anything that they pseudo-bear that pseudo-relation to. Such "objects," after all, can be nonexistent. Thomas Reid's "ambulo ergo sum" would be appropriate here, not the Cartesian *Cogito* conceived in a phenomenologically performative way.

This situation is rather paradoxical. If the only mode of awareness of our own consciousness (even supposing ubiquitous self-manifestation) is via phenomenal intentionality so construed, then our evidence for the very existence of our own consciousness is really no better than our evidence for the existence of phenomenal colors. Just as we might be persuaded that there really are no phenomenal colors, perhaps we could become persuaded that there is no such thing as phenomenal consciousness either. I regard this as absurd. It is like saying that perhaps we only think we think, or that perhaps it only appears to us that things appear to us. Consequently, consciousness must bear some evidentially relevant relation to itself and to its own being, other than the phenomenal

intentionality pseudo-relation it pseudo-bears to phenomenal colors.

Thinking of consciousness as "being-appeared-to-existingly" does not help here, since that applies to phenomenal colors and all other perceived pseudo-objects and pseudo-qualities as well. Any theorist committed to self-manifestation should not try to construe this as just a case of phenomenal intentionality as just described. From our self-consciousness we can conclude that we do exist, and this is not just because we know by inference or in some other way that we are the bearer of a property, as in Reid's *Ambulo*. We must be acquainted with our own existence—in the sense that every episode of consciousness, however individuated, is acquainted with its own existence. This applies to the subject-pole. What about the object-pole?

In the context of the theory of perceptual consciousness, I think it is a mistake to maintain that any view according to which one can always legitimately quantify over the "immediate objects of conscious awareness" is committed either to some form of direct realism (or perhaps a disjunctivist version thereof) or to old-fashioned sense-datum theory. Any plausible form of representationalism—fictionalist or realist, externalist or internalist, reductive or non-reductive, is, I'll argue, committed to such quantification, though this must be understood in a particular way. I am not, of course, saying that if we seem to consciously visually perceive a pink rat then we can infer that there exists a pink rat that we see. There is, however, something other than just the conscious state itself (*qua* whole) that we can legitimately, existentially quantify over.

Our conscious perception of differentiation (in unity) entails, even on a representationalist view, that there exists something of which we are aware, namely, at the least, differentiation (or contrast) itself. For example, suppose I hallucinate purple and pinkish smoke clouds arising from stereo speakers as "Fairies Wear Boots" comes on. Evidently I cannot conclude that those purple and pinkish clouds exist. Still, I maintain, we can conclude that there exists some differentiation or contrast of which we are aware. By hypothesis, we cannot say that the difference is that between the pink smoke cloud and the purple one, since they

do not exist. Differences between non-existent objects cannot be appealed to in order to make sense of real differences.³¹ But we are aware of some *real* and phenomenally manifest differentiation here. If we say no to that, we'd have to assume that reflection is simply inaccurate when it comes to such hallucinations; that we seem to have a differentiated experience when in fact there is no phenomenal difference at all. But if that itself is a phenomenal state, say a conscious reflection on an ongoing hallucination, we have the same problem all over again.

If the difference we are aware of is not and is not to be accounted for by a difference in the objects (since they do not exist), it must be a matter of the difference in the representations. Hence, albeit in an indirect manner and, as it were, under the guise of a difference in the pink and purple clouds, we must be aware of some differentiation inherent in the representational states themselves.³² If we reject disjunctivisms, then we ought to maintain that in every case of differentiated phenomenal awareness we are, in fact, acquainted with (and not merely representing) the differences inherent in our episodes of phenomenal consciousness. This is, at any rate, what I think is the most plausible account, even if the considerations just given don't absolutely clinch it. Again, it is not that there cannot be some sort of representationalist response.³³ It is, rather, that I regard the line I take to involve fewer epicycles.

We cannot make good sense of the appearance of a phenomenal difference without direct awareness of differentiation. But, by hypothesis,

in the case of hallucination it cannot be that we are aware of a real difference in the objects of representation. Moreover, it cannot be a difference in something that is *hidden* from conscious awareness—some difference in the externalist conditions determining the content of the representational states, for example—that we are aware of. The most plausible candidate, then, is that we are directly aware of (acquainted with) differentiation or modifications in consciousness itself (and hence the Transparency Intuition (see page 4) is, strictly speaking, false; we are indeed aware of features of consciousness itself even in so-called “first-order” awareness). This applies to both reductive and non-reductive forms of representationalism. If this line of thought is correct, representationalist theories really presuppose some sort of non-representationalist, acquaintance theory.

Implicit in the above discussion is something like this definition of acquaintance:

Acquaintance =_{Def} (1) the relation (R) the subject (s) of consciousness (i.e., the episode or stream itself) bears to the differentiated phenomenal manifold ($D\langle x_1, x_2, \dots, x_n \rangle$), such that (2) if $sR[D\langle x_1, x_2, \dots, x_n \rangle]$, then we may infer truly that $(\exists x)(sRx)$.

Of course, clause (2) can be taken as redundant, given the usual understanding of real relations and that the R of clause (1) is so taken. But in this context it is important to emphasize the point. The first clause is just an inner-ostensive phenomenological characterization that assumes that the relational appearances are indeed the reality; the second is a logico-ontological characterization. Importantly, we can “quantify in” here: If, in any concrete particular case, we stand in that relation to some phenomenally differentiated field, then we can truly infer that *there exists* something differentiated we stand in that relation to. However, it is in general *not* the represented (or intentional) objects that we are thus acquainted with. It is, rather, the common factor of all episodes of phenomenal consciousness, be they hallucinations, dreams, or the “perceptions” of brains in vats. This, again, is often precisely what is denied when one says

³¹ We could possibly hold that even if the property instances are not real, the universals represented are, and try to account for the difference in phenomenology in terms of those real differences. But this sort of view does not allow us to make sense of the concrete but hallucinatory representation of different particular instances of the different properties.

³² I have briefly made similar arguments in Williford (2013).

³³ In particular, a representationalist could say that the represented difference between the pink and purple clouds is just as hallucinatory as the clouds themselves. This is, in a sense, correct. However, representationalists hold (or ought to hold, anyway) that phenomenal differences *always* correlate with differences in the representations themselves (and only normally in the objects of representation). If there are phenomenal differences, there exist some differences inherent in consciousness that are not merely the objects of representation. What I am claiming is that we are acquainted with this differentiation under the guise of differences in objects represented. An adherent of the Transparency Intuition would deny this, of course. And I don't take these considerations to constitute a knock-down argument. (Thanks to an anonymous reviewer for bringing this up.)

that a state is one of representation as opposed to acquaintance. If it is true that *I represent A*, I cannot infer from this that *there is some X such that I represent X*. Adverbialisms and other forms of representationalism were, recall, developed precisely around this insight in order to overcome the problems of sense-datum and other relational theories of perception. Is the theory I am suggesting here a form of old-fashioned sense-datum theory?

Unfortunately I cannot give a short answer to that question and can't give all of the long answer here. This will have to suffice: (1) We can regard sensory qualia (or hyle) as being complex, relational properties of consciousness (and its concrete embodiment in brain processes); in fact, they could be something like irresolvable structural properties that appear simple precisely because they mark a limit of our sensory resolution. (2) In order to flesh this out, we must reject the Revelation Thesis—the thesis that acquaintance yields up all of the properties of sensory qualia. In particular, we can (and should) reject the idea that acquaintance tells us all of the categorial properties of sensory qualities. There is no good reason to believe that it does. Hence, they could fail to seem relational and yet still be relational. This is a solution to the “Grain Problem”—a problem arising from the fact that brain properties are “complex” and relational while sensory qualities (phenomenal colors, tones) do not seem to be. If we infer from the appearances then we cannot consistently hold that they are identical to brain properties. But we have no good reason for making that inference.³⁴ (3) It is not hard to understand why the sensory qualities would be integrated into a spatialized and “intentionally animated” grid that can serve as a “user interface” for us to deal with the external world, yielding a “transparent” manifold in Metzinger’s sense, a manifold we are built to systematically and automatically “see right through”—causing us to suffer from a sort of delusion of direct realism (see Williford et al. 2012; Williford 2013; Metzinger 2004, p. 163, and Revonsuo

2006). Finally, (4) appeals to the “Transparency Intuition” (in Tye’s sense of “transparency”) thus carry no serious weight. All the phenomenological data in question are accounted for by 1–3, and there are good independent lines of reasoning for each of these (that we do not have time to go into here).

I’ve argued that the notion of acquaintance, when interpreted in the rather minimal, phenomenological, and logico-ontological way proposed, is the proper notion for characterizing the relationship between consciousness and the differentiated but unified multimodal experiential manifold. Moreover, on the view proposed here, consciousness bears this same relation, generically understood, to itself.

If the episode of consciousness bears the relation to itself, then evidently there is something to which it bears that relation. But, non-trivially, we could not have the sort of direct evidence of its existence that we do have if consciousness were not self-acquainted—and acquainted with its own existence. And if the episode of consciousness bears the relation to the differentiated manifold that constitutes the surface that serves as its contact with a differentiated reality beyond it—i.e., if it bears it to a differentiated portion of itself—then there is something differentiated of which it is non-representationally aware. One is directly aware of the *difference* or *differentiation* even if one only, strictly speaking, *represents* what the things so differentiated happen to be or interprets them as being such and such (mental, physical, surfaces of objects, internal sense data, quotidian objects, etc.). In other words, I can see that red is not blue even if I do not know what colors are exactly, or if they are in physical space or only in a virtual space in my brain. One does not *merely* represent this difference or differentiation. One is acquainted with oneself and with the differentiation one contains. Of course, one is also acquainted with the apparently intrinsic properties that mark these internal differences, but again, this need not mean that the properties are in fact non-relational and simple. In fine, we are self-acquainted and acquainted with a differentiated manifold and thus, at some level, with real differences in the mind, the

³⁴ I’ve argued this is in a bit more detail in Williford (2013). For relevant background ideas see Williford (2005 and 2007). For a discussion of the Revelation Thesis see e.g., Stoljar (2006, ch. 11) and Goff (forthcoming).

world, or world-mind boundary.³⁵ The acquaintance relation consciousness bears to itself is, generically speaking, identical to the relation it bears to sensory qualia (or hyle)—which are taken here as ultimately just transient modifications in the unfolding embodiment of consciousness. It is important to understand that this does not imply that there is a special type of sensory quality (a “me-ish” quale) peculiar to consciousness. It is as diaphanous as G.E. Moore said. Remember that the acquaintance in an instance of acquaintance with phenomenal red is identical with the acquaintance in an instance of acquaintance with phenomenal C#, even though phenomenal red and C# are utterly heterogeneous.

One might reasonably ask for a more substantive definition or account of acquaintance. The definition given relies on phenomenology and logic and is otherwise quite empty. But this is as it should be, in my view. Any further account of the nature of acquaintance, of what the acquaintance relation *is*, will be the result of empirical inquiry and a well-supported *a posteriori* identification.

5 Self-acquaintance, subjective character, and individuation

Earlier I briefly noted that at the phenomenological level we should probably not construe subjective character fundamentally as a matter of “mineness” or a “sense of self” where the latter is thought of as a sense of oneself as an owner of experiences. It is not that I do not think this description contains a grain of truth; I do. The worry, though, is that if we go this route, we might come to the conclusion that subjective character involves acquaintance with a haecceity—Zahavihood and Gallagherhood once again. Here I want to consider the same issue from a more ontologically oriented point of view.

We are indeed individuated and aware of ourselves (something individuated). And we can be aware of ourselves *as* distinct individuals and owners. But this does not at all entail the doctrine of haecceities immediately present to con-

sciousness—for-me-ness or me-ishness as a special property that no one else can share. Rather, subjective character is a common form that all conscious states have; but having this form does not alone make something the individual it is, evidently. It may be that in virtue of which we can be *aware* of ourselves as individuals, but it is not that in virtue of which we *are* the individuals we are. Yes, there is a determinate individual (somehow construed) that is acquainted with itself. No, this does not necessarily mean that it is acquainted with that *in virtue of which* it is individuated. That could be whatever it is that individuates physical objects. Or, perhaps, nothing is metaphysically individuated *by* anything else. But it ought to be clear that simply in being aware of myself I need not be privy to anything non-trivial about my metaphysical individuation conditions.³⁶

You are aware of your consciousness as something individual. You are a self-aware individual, if you prefer. But this does not mean that your subjectivity consists in being directly aware of *what* individuates you or the very property in virtue of which you are the individual you are. Or, perhaps, one may be aware of this property or set of properties, but only in the guise of being an individual that is thus and so. The “thus and so” part (all your contingent properties, your “facticity”) is radically changeable. You need not have been thus and so. (You could have been a contender! And if only you’d been rich!) You can also be aware that you are a particular instance. So, yes, you can become aware of your particularity. But everybody is aware of their own particularity. And it is, in a way, an empty and non-material (in the “formal vs. material” sense) property. It’s not as if my particularity has a special something that yours lacks and vice versa. Hence, I would not be able to tell, by phenomenological intuition alone (or in any other way for that matter), which of the infinitely many duplicate and near-duplicate worlds I am in (cf. Elga 2004). Am I in the world in which one of Napoleon’s buttons had a bit of his blood on them the morning of the Battle of Jena or in the world in which that was

³⁵ I have considered our acquaintance with a differentiated manifold *qua* mind-world boundary in more detail in Williford (2013).

³⁶ I have briefly argued this before in Williford (2011b). I was pleased to find that a similar line of argument was pursued by the eleventh-century Buddhist philosopher Ratnakīrti; see Ganeri (2012, p. 217).

not the case? I cannot tell by introspection, yet, depending on the correct answer, I am one type of individual (and of course, one token of uncountably many of that type) and not of the other type (which type also contains uncountably many individual counterparts of mine). I am individuated, and I know that; I belong to just one of these worlds. But I do not have complete access to my individuation conditions or the conditions, if there are any, that determine that this individual is in one world as opposed to another. I have uncountably many counterparts who feel exactly the same way because, to speak loosely, they don't know that they are not me; none of us can tell the difference. I cannot locate my Homeworld on the map of worlds that contains my relevant counterparts.

It is a mistake, then, to make subjective character depend on the sense of individuality; this reverses the proper order of explanation. Self-acquaintance and concrete instantiation yield the sense of individuality, and they do it again and again in many places and in the same way. Evidently, the contingent filling that experience and history infuse into the formal shell of conscious subjectivity is not relevant at the level we are concerned with. Hence, it can also be metaphysically, not just phenomenologically, misleading to use terms like “for-me-ness,” “mineness,” “me-ishness,” etc. That is to make something derived seem like something basic. The basic things are self-acquaintance (“reflexivity”) and actual, concrete instantiation or constitution. The sense of individuality comes from these, not the other way around.

Of course, if you are a real, concrete individual, you are individuated. But individuation is evidently not self-acquaintance. The latter is, however, required if one is to get the *sense* of being an individual, to know, feel, and be concerned with oneself *as* an individual. If we generally equated self-acquaintance with something's being the individual it is, then we'd have to hold either that every individuated thing in the cosmos is self-acquainted and conscious, or that conscious things have one type of metaphysical individuation conditions, and non-conscious things another, for very obscure reasons. Moreover, we either must not take subjective character to be a

univocal notion or must resort to some sort of hopeful brute resemblance nominalism about subjective character and maintain that we cannot not really know that, say, I, *qua* subject, am in any meaningful sense like you, *qua* subject. This is not a very good dilemma to be in.³⁷ I think the more plausible view is that self-acquaintance is not the source of the individuation of consciousness but rather something that both concretely depends upon individuation and enables the knowledge of individuality and, consequently, self-location in surrounding spaces.

It is misleading, then, both phenomenologically and ontologically to refer to subjective character principally as “mineness” or “me-ishness” or “for-me-ness,” even though subjective character is one of the bases of the sense of individuality. We should not think of self-acquaintance (and subjective character) as anything more than this relation all episodes of consciousness bear to themselves. It is a perfectly uniform structure and a kind of universal—in that sense, supposing one is some sort of realist about universals, there is indeed some identical thing that unifies all episodes (or subjects) of consciousness, namely the very property of being self-manifesting; but we are all distinct instances. Thus, in a very special and non-Vedantist sense, we could say that there are many instances of consciousness but only one subject, with some instances connected to each other and grouped together in other important ways as well. But there is no substantial self. In this regard, I am with Hume, Sartre, Parfit, Strawson, Metzinger, the Buddhists, and other “non-egological” theorists of consciousness. Note that this does not mean that consciousness is “anonymous” in the sense of “subjectless.” Every stream of consciousness has its transient subject (*viz.*, itself) but that is not a substantial self.

6 Self-acquaintance, intrinsic properties, and physicalism

Should we really regard self-acquaintance as a relational matter? Is it really a matter of some

³⁷ Previous episodes of consciousness normally connected to the present episodes (the ones producing this document) found themselves trying to live with the latter horn of the dilemma in the flawed Williford (2005).

sort of thing standing in a relation to itself? On the one hand, there is no special problem either logically or phenomenologically speaking with the idea of something relating to itself in this way. Appearance is appearance-to. That's relational. There is no *a priori* reason why something could not appear to itself. It does not lead to a regress.³⁸ One should put aside misleading and question-begging spatial analogies—consciousness is not like a knife trying to cut itself. Advocates of self-acquaintance will claim, opposing one analogy with another, that it is more like a candle's flame illuminating itself by emission while it illuminates other things by the reflection of its light; it does not require another candle flame for it to be illumined.³⁹ Moreover, one must remember to exclude from one's mind the sort of objectification and description-based cognition that normally overlays the phenomenal manifold. We are talking about the sphere of immanence, to speak Husserlian, and not about intentional objects or constituted objectivities given via *Abschattungen*. Again, we are talking about immediate self-acquaintance, not the representation of oneself as being such and such. It is indeed more like the emission of light than the reflection of light, if we must pick an analogy.

Nevertheless, even if we accept the relational construal and remember that it is an immediate and direct relation not mediated by concepts or descriptions, we still have a problem. It is not as if conscious episodes just happen to be self-manifesting. The property of being self-manifesting is not something that a thing can have and then not have—like changing coats of paint. It is of the very essence of a conscious episode. This is not an external relation to itself or one mediated by convention or history or anything else. Hence, it must have some set of intrinsic properties in virtue of which it is self-manifesting. Thus, the Heidelberg School, Michel Henry, and Dan Zahavi, I'll

concede, win on this ontological point. Dieter Henrich, Manfred Frank, Henry, and Zahavi have all maintained that self-manifestation could not be a relational matter (e.g., [Henrich 1971, 1982](#); [Frank 2002, 2007](#); [Zahavi 1999](#); [Henry 1973](#)). And they are very close to being right. I think, however, that it is more accurate to say that even if it is a relational matter, it is not an *external* relation we are dealing with. So there must be something about the internal structure of consciousness that grounds the relation. In short, as Henrich and Frank have long said, there must be some intrinsic property in virtue of which episodes of consciousness (out of all other things in the world) are self-manifesting. What could this property be? Are we left with something that cannot be physical, or, even if it is physical, is nevertheless irreducible in some sense?

It may seem now that David Rosenthal is having his revenge.⁴⁰ In effect, I have been arguing against the extrinsicalist view—the view that something's being conscious has to do with external relations the thing stands in—be those external relations to other mental states or external relations to historically distant states of affairs or to other parts of one's cognitive apparatus. Now, to our chagrin, it seems we are left with something explanatorily basic. At this point we are left with two problematic strategies. We could go the panpsychist route ([Strawson 2006](#)): It's no surprise that we're conscious if everything is! Or if, as I do, one thinks (after Locke in a similar context) that “every sleepy nod doth refute” this, we can hold that only certain physical complexes instantiate this particular property (or set of properties). This will mean either some form of property dualism or some form of identity theory (possibly with its “Harder Problem”; see [Block 2002](#)). If one does not want to be a dualist or a panpsychist, what can be said?

Here is the sort of approach that seems most attractive (to me, anyway). We want to hold that consciousness is indeed some sort of physical process. It's not, however, just a matter of the satisfaction of some functional role. I

³⁸ This is demonstrable. First, obviously, there is no logical problem with reflexive relations. Second, it requires special and highly questionable premises to generate another regress here. See [Williford \(2006\)](#). See also [Kriegel \(2009, p. 124\)](#) and [Janzen \(2008, p. 110\)](#).

³⁹ The knife blade and candle flame competing analogies loom large in the Indo-Tibetan debate on this issue. Clearly, the analogies will be found, by opponents and proponents, to be exactly as plausible as the views they encode.

⁴⁰ Though even Rosenthal's own view was pushed into being (or always was) problematic in this regard, as noted above.

think it also has a functional role. But it is not *in virtue of* playing that role that something is consciousness; rather, consciousness is suitable for that role because of its properties.⁴¹ In principle, many different things could play that role (at least if we specify it entirely in behavioral terms). Or, at least, this is an open question. Consciousness has a functional role, but it is not to be identified with just any arrangement of elements that can play that role as causally and behaviorally specified. There is some special, distinctive physical process that is consciousness. It plays its functional role in virtue of its having the properties it does and *not vice versa*. But then does some version of Russellian Monism start to seem attractive (see e.g., [Stoljar 2006](#), ch. 6 and [Pereboom 2011](#), chs. 5 and 6)? Am I saying that the functional role is just being (contingently) satisfied by a (somehow) unified and self-manifesting group of qualia? Or something *wild* like that?

Here we play the same sort of trick we played when dealing with the Grain Problem. Consciousness is self-acquainted, but we are also, as Fumerton and Fales would say, acquainted with acquaintance; we are given givenness ([Fumerton 1985](#), pp. 57–58 and [Fales 1996](#), pp. 147–148). The relation does not seem complex or to involve many layers of relational structures. But we cannot infer from this appearance that it is in fact such a simple relation. Again, its not seeming complex does not, without controversial and implausible completeness assumptions, entail its being simple. Moreover, once we realize that normal consciousness involves a great many intricately related aspects—at least (non-contingently) differentiated unity and temporality, and (contingently) animation functions operating on a differentiated sensory manifold, iterations of these functions, pattern extractions, etc.—we have all the more reason to suppose that there is complicated machinery hidden from our introspective view. In fact, it will be noted in a Sartrean and Moorean vein, that consciousness, both as acquaintance relation and subject-relatum, seems mightily empty. Once we realize that Revelation theses fail, then we no longer need read this ap-

pearance as “consciousness *qua* acquaintance relation appears simple.” Rather, we read it as “consciousness *qua* acquaintance relation *does not appear* complex.” These are, in many cases, phenomenologically indistinguishable, but they are logically different.⁴² The first reading, coupled with an infallibility thesis (or with just a strong presumption in favor of the deliverances of naïve introspection), leads to the view that acquaintance is simple. But the other requires a Revelation (or completeness) thesis to get the same result. Revelation is, again, totally implausible. And even if we were to assume infallibility, we have no a priori reason to favor one interpretation of the phenomenological data over the other—the “seeming non-P” vs. “not seeming P” formulation. We do, however, have plenty of *a posteriori* inductive reasons for preferring the latter: It does not seem complex, but it is (or at least could be for all we can tell phenomenologically).

Since we have an extremely limited resolution when it comes to penetrating into the nature of consciousness by introspective means, we are quite free to adopt another strategy. We can accept an *a posteriori* identity theory. Consciousness is identical to some sort of recurrent physical process unfolding in the brain. Fundamentally, what we get from introspection is a sort of structure and some irresolvables—the sensory qualities—that are like reflections of the materials in which the form or structure is instantiated. Since we have rejected Revelation (completeness) theses, we can accept that sensory qualities (and the acquaintance relation itself) are complex and involve layers of relations even though they do not seem this way (just as the headless woman⁴³ in the famous illusion does not seem to have a head—absence of appearance is transformed into the appearance of an absence; see [Armstrong 1968](#) and [1973](#)).

⁴² They are phenomenologically indistinguishable in the way that the stream of consciousness’s being temporally continuous is, plausibly, phenomenologically indistinguishable from consciousness’s being punctate or discrete, or in the way in which consciousness’s seeming free from causal determination is phenomenologically indistinguishable from its simply not seeming determined (because the causal relations are inaccessible, as Spinoza suggested).

⁴³ See the following links:
<https://www.youtube.com/watch?v=LXOqD5B5Sxc>
<http://www.deceptology.com/2010/10/headless-woman-illusion.html>

⁴¹ Here I am in considerable agreement with [Langsam \(2011, ch. 3\)](#).

We can use what structure we are aware of, however, to build models to guide our search for the neural correlates of consciousness. One thing we see is that the (only apparently simple) acquaintance relation involved is such that whenever xRy , xRx ; while it is not the case that if xRy , then yRy (in the case where y is a sensory quality or manifold thereof). And we have some idea of what the qualities in the manifold could be—e.g., limits of resolution or irresolvables operated on by a spatializing filter. We can also see that spatial projection, integration of multimodal information, temporality, and the modulation of attention are involved (along, of course, with more advanced things like intentional animation, cognitive filtering and reprocessing, and poise for action). We have a self-manifesting totality containing a unified and spatialized but differentiated manifold. Consequently, we do need to look for processes that can do information integration and binding, but that is only necessary, not sufficient. We need to look as well at processes that spatialize the multimodal (and multidimensional) information (see [Williford et al. 2012](#)).

This does not at all mean we are looking for a little room in the brain that has patches of red, yellow, blue, and green mental paint in it. Rather we must look for more abstract correspondences. In the case of the sensory qualities, we are possibly looking for higher-order relations between fairly complex structures, structures that can transiently be pulled into and “rendered” by the core process. Basically, this panoply of contrast-related irresolvables gets generated in a real-time and transient fashion, now occupying this virtual “location”, now occupying that, depending on a whole host of input factors (head orientation, background, conceptualization, etc.). These “locations” map onto (we hope) real physical space at a certain scale, but it is not a matter of finding a “bubble within a bubble.” It is a matter of an abstract correlation of structure. The isomorphisms (or homomorphisms) could be there even if the internal “space” of experience is entirely virtual, a kind of computational “movie in the brain” to use another phrase of Damasio’s. Assuming the principle that the positive and critically evalu-

ated set of phenomenological descriptions gives us not just the way consciousness seems, but the way it in fact is, along with our identity postulate, we can be sure that something in the brain has a structure corresponding to this, no matter how transformed by “layers of abstraction” it may be.

What is more, self-acquaintance will demand that we explore models in which real reflexivity can be encoded. Hofstadter’s model is one of these.⁴⁴ But following D. Rudrauf and further encouraged by D. Bennequin, I have moved in the direction of considering projective geometrical models. There is no space to go into this here, but suffice it to say that there exist mathematical frameworks that allow us to conceptualize and investigate more deeply the self-acquaintance-related features of consciousness by considering the interplay of the space we project and the origin of the projection (see [Williford et al. 2012](#) and [Rudrauf et al. ms](#)).

The goal of such work would be the refinement of mathematical models of the structure of consciousness. Upon the achievement of that end, we would then try to determine how such models could possibly be physically realized in the brain. Once we can say what the physically detectable signatures of such a realization might be, then we could one day meaningfully test such theories. Were we to verify the existence of such a structured process in the brain, explaining consciousness would reduce to explaining how the process is realized—what parts have to be in what order doing what and at what time scale.

It will always seem to be a brute fact, at some level, that consciousness is physical process X , however X gets fleshed out. But we’ll just get used to it, as long as there is some somewhat intelligible bridge (in this case provided by mathematical models) from the lived phenomenon to its brain correlate. We’ll get used to it just as we’ve gotten used to water being H_2O . It could be that there will be multiple ways to implement such a process. Sup-

⁴⁴ While Hofstadter’s Gödel-inspired model might be problematic (both in terms of physical implementation and in terms of the strong mathematical realism it might presuppose), it is certainly in the right class of models we should be considering. See [Hofstadter \(2007\)](#).

pose, just for example, that it has to do with generating certain types of fields and that multiple substrates, not just brains, can generate and support the relevant sorts of fields. Then consciousness will be, to that degree, multiply realizable. Suppose it is a matter of realizing a certain computational organization. Then, in effect, implementing a certain program will be equivalent to being conscious; and if machines made from different substrates or with different architectures can run the program, consciousness will be multiply realizable in the sense of computational functionalism. Your particular consciousness then, as you know and love it, would be just the concrete running of the program in your particular brain.

We might wonder, in such a case, what it is to “run” a program or to “have” a certain structure or to “instantiate” such an arrangement or system of fields or whatever. Of course, this is a quite general metaphysical problem that we should not confuse with any problem specific to consciousness. However, given that we are acquainted with our own individual existence, it seems that somehow its instantiation makes its very instantiation available or manifest in some non-representational way. This is rather peculiar. If we are going to be physicalists who are nonetheless responsible to the phenomenology, however, this is what we have to accept, or so I have argued. Something is conscious if it has a certain internal structure and attendant dynamical profile. Being conscious *is* having that structure and profile. We will never be able to explain *why* that is the case because it is simply a confusion to think that identities like this admit of explanation; they can only be discovered (Papineau 2002, ch. 3). We must, of course, give evidence in favor of the relevant identity claim; uncovering such evidence is the goal of scientific research on consciousness. Our choice is between this sort of view and the view that *there is something else, something non-physical* that just *is* consciousness. Of course, we’d never be able to explain why *that* is the case either. So in the absence of compelling arguments for dualism or panpsychism, Occam’s Razor would lead us, as Smart pointed out so long ago, to embrace an identity theory.

The identity theory only adumbrated here would be neither a crude type-type identity theory nor a causal-role functionalist token-token identity theory where the realizers do not matter at all. Since any concrete consciousness is a marriage of form and matter (and the self-appearance of that marriage), and since there no doubt are physical constraints on what sorts of materials can be put into that form, we want to identify consciousness with neither a specific type of material (or “wonder tissue” in Dennett’s phrase) nor with an abstract, disembodied form that seems trivially realizable by practically any set of elements—since purely abstract isomorphisms may be a dime a dozen.⁴⁵ In other words, we need a non-eliminativist and non-idealist account of what it is to *really* realize a structure, instantiate a form, or, as the case may be, to *really* run a program or compute a function. To my knowledge, no one currently has such an account.

At bottom, this is just the old metaphysical problem of the Methexis—the relation of universals to particulars or of form to matter. When I am feeling optimistic, I imagine that I’ve reduced the problem of consciousness to another, more general (as well as ancient and probably insoluble) metaphysical problem. We may not know what it is for matter to really and mind-independently take on a certain form, but it is hardly an implausible metaphysics that says that this happens. It is arguably this type of metaphysical view that would best explain the success of applied mathematics, engineering, and the sciences: they are successful because the world really does have (or approximate) the relevant mathematical structures—these are *in re* structural universals. This seems to be a commitment of scientific realism. But perhaps we will never get beyond a rather crude operationalism when we empirically investigate such matters; perhaps the metaphysical nature of property instantiation will forever remain obscure to us. That should not, however, discourage us from carrying on such empirical investigations in the case of consciousness. Even if there will be

⁴⁵ For discussion, see Chalmers (1996) and Buechner (2008, ch. 3).

a residual metaphysical mystery, it is a general one, not one specific to consciousness.

The main point here, and the concluding one, is that consciousness could be self-acquainted, where this is not a matter of external relations, and still some form of relatively non-mysterious (hylomorphic) physicalism could be true. One might balk at the idea that this would not be a matter of external relations, especially if we go the computational functionalist route. But think of it like this: If we are realists about the implementation of computational structures, then even though the structures involve parts and elements, there is still a unity to the pattern as implemented. It is, in a certain sense, an indivisible whole that is not just the mereological sum of its parts. Analogously, the circle has its own structure and characteristic properties even though it is made of points. What we really need, and may never have (but who knows?) is a theory that tells us when we have a real, concrete unified whole, (where this is *not* simply a functional or conventional characterization but is a matter of more basic physical relations) and when we have unities and wholes (and instantiations of structures and properties) that are only conventionally real.

Suppose then that we adopt a sort of realism about computational (or otherwise structural) wholes, which we have some independent reason to do. Circles have remarkable properties, *qua* circles, even if they are made up of points. Concrete circular things approximate these. Simultaneous cycles have certain number-theoretic properties just *qua* cycles regardless of *what* they are cycles of (e.g., reproducing cicadas and cicada predators, see [Baker 2005](#)). Likewise, for the concrete implementation of consciousness, it is surely the case that certain elements must be put into a certain arrangement, realizing a certain structure and dynamics. This would not mean, however, that consciousness *as such* is to be identified with either those elements or the arrangement abstractly conceived. Rather it is the *concretely implemented* organization of those elements *qua* whole. In virtue of being an instance of that form or structure, it has

certain properties. One of these could be the property of being self-manifesting. That property could itself be a complex relational property having a certain unity. The account sketched here presupposes a certain realism about the instantiations of mathematical and computational structures—that there are determinate, mind-independent facts of the matter about this. We cannot go further into this rather large and complicated metaphysical hornet’s nest. Suffice it to say that a real, unified, concretely instantiated structure could, in a certain sense, be relational and have components even if it is, in another sense, an intrinsic property.

7 Conclusion

I have argued that the best way to characterize subjective character is in terms of self-acquaintance and not, for various reasons, in terms of Higher-Order, Same-Order, or Privileged-Object representation. I argued that every episode or stream of consciousness is acquainted with itself, and not with a self in some other sense—a homunculus, substance, or haecceity. This is, I maintain, the best way to make sense of the intuition of subject-object polarity and the Humean intuition that we do not find a self-entity. Moreover, one’s sense of being an individual is a consequence of self-acquaintance and concrete existence and not to be conflated with subjective character as such. Such conflation leads to potentially misleading descriptions of subjective character (as “mineness”) and, if taken literally, to metaphysically and epistemologically undesirable consequences. We are individuated and self-acquainted, and that is enough to allow us to derive the sense of self or “mineness”; but self-acquaintance is not itself what individuates us, nor does it necessarily make us aware of what does.

Nevertheless, I conceded to Henrich, Frank, Henry, and Zahavi (among others) that consciousness must have some intrinsic (or internal relational) property in virtue of which it is self-acquainted. But I argued that this does not nullify the appropriateness of de-

scribing subjective character as being a matter of a very complex relation, though it does not seem to be so complex.

Finally, I argued that the position advanced here is not incompatible with a form of (hylomorphic) physicalism. Sensory hyle, the acquaintance relation itself, the self-manifesting episodes, could all be brain processes and properties. On the phenomenological side, this gains plausibility once we take to heart the incompleteness of introspection (and of pre-reflective self-awareness as well): not seeming complex and relational does not entail not being complex and relational. On the ontological side, I argued that even some form of computational functionalism could be true. But, generally, the important thing to remember is that consciousness is the marriage of form and matter. It cannot be simply equated with either. This opens up space for multiple realizability, but it might also mean that not just any old substrate will do. It's an open question. The metaphysical commitment behind this position is just some form of realism about structural universals and their mind-independent instantiation conditions, which is arguably a commitment of scientific realism in any case. Absent dualism, panpsychism, or idealism, that is what we will have to accept, I believe. (Eliminativism is, of course, a non-starter.)

We do not need a theory of the Methexis, however, in order to attempt to find the neural correlates (correlation conceived of as indicating identity here) of consciousness by building mathematical models of the phenomenology and figuring out how the brain might implement the structures so modeled. In fact, just such an approach is quite in line with scientific practice generally: We know that the world we investigate with our relatively crude means is, in multiple ways, a play of matter and form even if we do not really know what the Matter ultimately is, what Forms are, and how the latter come to live in the former.

Acknowledgements

Different parts of this material were presented at many places over several years. I would like

to thank audiences at the Johannes Gutenberg-Universität Mainz (seventeenth meeting of the MIND Group), the Berlin School of Mind and Brain, the Institut Jean Nicod, ZiF, SMU, the SSPP, TCU, and Tucson TSC for relevant discussions. I would like to thank these institutions and the symposia organizers (and in particular, Thomas Metzinger and Jennifer Windt; Manfred Frank, Marc Borner, Andreas Heinz and Anna Strasser; Brad Thompson and Philippe Chuard; Pete Mandik, Rik Hine, and Blake Hestir; and David Chalmers). Thanks to the College of Liberal Arts and the Department of Philosophy and Humanities at the University of Texas-Arlington for research and travel funding in this connection. I should thank (in alphabetical order) Katalin Balog, Daniel Bennequin, Jacob Berger, Alexandre Billon, Marc Borner, Philippe Chuard, Christian Coseru, Justin Fisher, Manfred Frank, Brie Gertler, Robert Howell, Tomis Kapitan, Bob Kentridge, Chad Kidd, Alex Kiefer, Uriah Kriegel, Greg Landini, Stefan Lang, Pete Mandik, Thomas Metzinger, Charles Nussbaum, David Papineau, Gerhard Preyer, Harry Reeder, David Rosenthal, Amber Ross, David Rudrauf, Susan Schneider, Miguel Sebastian, Charles Siewert, Anna Strasser, Brad Thompson, Keith Turausky, Michael Tye, Josh Weisberg, and Dan Zahavi for discussions, questions, criticisms, suggestions, etc., that were in one way or another of help to me in relation to the material presented here. In the same regard, I should thank two anonymous reviewers from the MIND Group for helpful feedback on an earlier version of this article; their feedback helped me to see some of my less-than-admirable tendencies as a writer of philosophy, even if it did not enable me to correct all their manifestations. Special thanks to Ying-Tung Lin of the MIND Group for her help. Special thanks to Trish Mann, Swathi Prabhu, Emma Nwokonko, and Anya Williford for help with the references. And very special thanks, once again, to Thomas Metzinger and Jennifer Windt for launching and managing this unique and ambitious project and to the Barbara-Wengeler-Stiftung for its support.

References

- Armstrong, D. (1968). The headless woman illusion and the defence of materialism. *Analysis*, 29 (2), 48-49. [10.1093/analys/29.2.48](https://doi.org/10.1093/analys/29.2.48)
- (1973). Epistemological foundations for a materialist theory of the mind. *Philosophy of Science*, 40 (2), 178-193. [10.1086/288514](https://doi.org/10.1086/288514)
- Baars, B. (1997). *In the theater of consciousness: The workspace of the mind*. New York, NY: Oxford University Press.
- Baker, A. (2005). Are there genuine mathematical explanations of physical phenomena? *Mind*, 114 (454), 223-238. [10.1093/mind/fzi223](https://doi.org/10.1093/mind/fzi223)
- Balog, K. (2012). Acquaintance and the mind-body problem. In C. Hill & S. Gozzano (Eds.) *New perspectives on type identity: The mental and the physical* (pp. 16-42). Cambridge, UK: Cambridge University Press.
- Berger, J. (2013). Consciousness is not a property of states: A reply to Wilberg. *Philosophical Psychology* (ahead-of-print), 1-14. [10.1080/09515089.2013.771241](https://doi.org/10.1080/09515089.2013.771241)
- Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18 (2), 227-247. [10.1017/S0140525X00038188](https://doi.org/10.1017/S0140525X00038188)
- (2002). The harder problem of consciousness. *The Journal of Philosophy*, 99 (8), 391-425. [10.2307/3655621](https://doi.org/10.2307/3655621)
- (2011). The higher order approach to consciousness is defunct. *Analysis*, 71 (3), 419-431. [10.1093/analysis/anr037](https://doi.org/10.1093/analysis/anr037)
- Buechner, J. (2008). *Gödel, Putnam, and functionalism: A new reading of representation and reality*. Cambridge, MA: MIT Press.
- Butchvarov, P. (1979). *Being qua being: A theory of identity, existence, and predication*. Bloomington, IN: Indiana University Press.
- (1998). *Skepticism about the external world*. Oxford, UK: Oxford University Press.
- Byrne, A. (2001). Intentionalism defended. *Philosophical Review*, 110 (2), 199-240. [10.1215/00318108-110-2-199](https://doi.org/10.1215/00318108-110-2-199)
- Cappelen, H. & Dever, J. (2013). *The inessential indexical*. Oxford, UK: Oxford University Press.
- Carruthers, P. (2000). *Phenomenal consciousness: A naturalistic theory*. Cambridge, UK: Cambridge University Press.
- (2005). *Consciousness: Essays from a Higher-Order Perspective*. Oxford, UK: Oxford University Press.
- Caston, V. (2002). Aristotle on consciousness. *Mind*, 111 (444), 751-815. [10.1093/mind/111.444.751](https://doi.org/10.1093/mind/111.444.751)
- Chalmers, D. (1996). Does a rock implement every finite-state automaton? *Synthese*, 108 (3), 309-333. [10.1007/BF00413692](https://doi.org/10.1007/BF00413692)
- (2003). The content and epistemology of phenomenal belief. In Q. Smith & A. Jokić (Eds.) *Consciousness: New philosophical perspectives* (pp. 220-272). Oxford, UK: Oxford University Press.
- Clark, A. (1989). The particulate instantiation of homogeneous pink. *Synthese*, 80 (2), 277-304. [10.1007/BF00869488](https://doi.org/10.1007/BF00869488)
- Coseru, C. (2012). *Perceiving reality: Consciousness, intentionality, and cognition in buddhist philosophy*. Oxford, UK: Oxford University Press.
- Crowell, S. (2011). Idealities of nature: Jan Patočka on reflection and the three movements of human life. In I. Chvatík & E. Abrams (Eds.) *Jan Patočka and the heritage of phenomenology* (pp. 7-22). Berlin, GER: Springer.
- Dainton, B. (2000). *Stream of consciousness: Unity and continuity in conscious experience*. London, UK: Routledge.
- (2008). *The phenomenal self*. Oxford, UK: Oxford University Press.
- Damasio, A. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. New York, NY: Harcourt.
- (2010). *Self comes to mind*. New York, NY: Pantheon Books.
- Dehaene, S. & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, 79 (1), 1-37. [10.1016/S0010-0277\(00\)00123-2](https://doi.org/10.1016/S0010-0277(00)00123-2)
- Dretske, F. (1995). *Naturalizing the mind*. Cambridge, MA: MIT Press.
- Dreyfus, G. (2011). Self and subjectivity: A middle way approach. In M. Siderits, E. Thompson & D. Zahavi (Eds.) *Self, no self?: Perspectives from analytical, phenomenological, and Indian traditions* (pp. 114-156). Oxford, UK: Oxford University Press.
- Edelman, G. & Tononi, G. (2000). *A universe of consciousness: How matter becomes imagination*. New York, NY: Basic Books.
- Elga, A. (2004). Defeating Dr. Evil with self-locating belief. *Philosophy and Phenomenological Research*, 69 (2), 383-396. [10.1111/j.1933-1592.2004.tb00400.x](https://doi.org/10.1111/j.1933-1592.2004.tb00400.x)
- Fales, E. (1996). *A defense of the given*. New York, NY: Rowman & Littlefield.
- Frank, M. (2002). Self-consciousness and self-knowledge: On some difficulties with the reduction of subjectivity. *Constellations*, 9 (3), 390-408. [10.1111/1467-8675.00289](https://doi.org/10.1111/1467-8675.00289)

- (2004). Fragments of a history of the theory of self-consciousness from Kant to Kierkegaard. *Critical Horizons*, 5 (1), 53-136. [10.1163/1568516042653567](https://doi.org/10.1163/1568516042653567)
- (2007). Non-objectal subjectivity. *Journal of Consciousness Studies*, 14 (5-6), 5-6.
- Fumerton, R. (1985). *Metaphysical and epistemological problems of perception*. Lincoln, NE: University of Nebraska Press.
- Ganeri, J. (2012). *The self: Naturalism, consciousness, and the first-person stance*. Oxford, UK: Oxford University Press.
- Gennaro, R. (2005). The HOT theory of consciousness: Between a rock and a hard place? *Journal of Consciousness Studies*, 12 (2), 3-21.
- (2006). Between pure self-referentialism and the (extrinsic) HOT theory of consciousness. In U. Kriegel & K. Williford (Eds.) *Self-representational approaches to consciousness* (pp. 221-248). Cambridge, MA: MIT Press.
- (2012). *The consciousness paradox: Consciousness, concepts, and higher-order thoughts*. Cambridge, MA: MIT Press.
- Gertler, B. (2011). *Self-knowledge*. London, UK: Routledge.
- (2012). Renewed acquaintance. In D. Smithies & D. Stoljar (Eds.) *Introspection and consciousness* (pp. 93-128). Oxford, UK: Oxford University Press.
- Goff, P. (forthcoming). Real acquaintance and physicalism. In P. Coates & S. Coleman (Eds.) *Phenomenal qualities: Sense, perception and consciousness*. Oxford, UK: Oxford University Press.
- Goldman, A. (1993). Consciousness, folk psychology, and cognitive science. *Consciousness and Cognition*, 2 (4), 364-382. [10.1006/ccog.1993.1030](https://doi.org/10.1006/ccog.1993.1030)
- Harman, G. (1990). The intrinsic quality of experience. *Philosophical Perspectives*, 4, 31-52. [10.2307/2214186](https://doi.org/10.2307/2214186)
- Henrich, D. (1971). Self-consciousness, a critical introduction to a theory. *Man and World*, 4 (1), 3-28. [10.1007/BF01248576](https://doi.org/10.1007/BF01248576)
- (1982). Fichte's original insight. In D. Christensen (Ed.) *Contemporary German philosophy: Volume 1* (pp. 15-53). University Park, PA: Penn State University Press.
- Henry, M. (1973). *The essence of manifestation*. Den Haag, NL: Nijhoff.
- Higginbotham, J. (2003). Remembering, imagining, and the first person. In A. Barber (Ed.) *The epistemology of language* (pp. 496-533). Oxford, UK: Oxford University Press.
- (2010). On words and thoughts about oneself. In F. Recanati, I. Stojanovic & N. Villanueva (Eds.) *Context-dependence, perspective and relativity* (pp. 253-282). Berlin, GER: De Gruyter.
- Hofstadter, D. (2007). *I am a strange loop*. New York, NY: Basic Books.
- Hopp, W. (2011). *Perception and knowledge: A phenomenological account*. Cambridge, UK: Cambridge University Press.
- Horgan, T. & Kriegel, U. (2007). Phenomenal epistemology: What is consciousness that we may know it so well? *Philosophical Issues*, 17 (1), 123-144. [10.1111/j.1533-6077.2007.00126.x](https://doi.org/10.1111/j.1533-6077.2007.00126.x)
- Howell, R. (2013). *Consciousness and the limits of objectivity: The case for subjective physicalism*. Oxford, UK: Oxford University Press.
- Janzen, G. (2008). *The reflexive nature of consciousness*. Amsterdam, Netherlands: John Benjamins.
- Kapitan, T. (2006). Indexicality and self-awareness. In U. Kriegel & K. Williford (Eds.) *Self-representational approaches to consciousness* (pp. 379-408). Cambridge, MA: MIT Press.
- Kidd, C. (ms). The idols of inner-awareness: Towards disjunctive self-representationalism. 5th Online Consciousness Conference.
- Kiefer, A. (2012). Higher-order representation without representation. 104th Annual Meeting of the Southern Society for Philosophy and Psychology in Savannah, GA.
- Kriegel, U. (2006). The same-order monitoring theory of consciousness. In U. Kriegel & K. Williford (Eds.) *Self-representational approaches to consciousness* (pp. 143-170). Cambridge, MA: MIT Press.
- (2007). The phenomenologically manifest. *Phenomenology and the Cognitive Sciences*, 6 (1-2), 115-136. [10.1007/s11097-006-9029-8](https://doi.org/10.1007/s11097-006-9029-8)
- (2009). *Subjective consciousness: A self-representational theory*. Oxford, UK: Oxford University Press.
- (2011). *The sources of intentionality*. Oxford, UK: Oxford University Press.
- (Ed.) (2013). *Phenomenal intentionality*. Oxford, UK: Oxford University Press.
- Kriegel, U. & Williford, K. (Eds.) (2006). *Self-representational approaches to consciousness*. Cambridge, MA: MIT Press.
- Lane, T. & Liang, C. (2011). Self-consciousness and immunity. *Journal of Philosophy*, 108 (2), 78-99.
- Langsam, H. (2011). *The wonder of consciousness*. Cambridge, MA: MIT Press.

- Levine, J. (2001). *Purple haze: The puzzle of consciousness*. Oxford, UK: Oxford University Press.
- Lockwood, M. (1993). The grain problem. In H. Robinson (Ed.) *Objections to physicalism* (pp. 271-292). Oxford, UK: Oxford University Press.
- Lycan, W. (1996). *Consciousness and experience*. Cambridge, MA: MIT Press.
- Lycan, W. (ms). A simple point about an alleged objection to higher-order theories of consciousness. <http://www.unc.edu/~ujanel/A%20Simple%20Point%20about%20an%20Alleged%20Objection%20to.pdf>
- Macphail, E. (1998). *The evolution of consciousness*. Oxford, UK: Oxford University Press.
- Mandik, P. (2009). Beware of the unicorn: Consciousness as being represented and other things that don't exist. *Journal of Consciousness Studies*, 16 (1), 5-36.
- (forthcoming). Conscious-state anti-realism. In C. Munoz-Suarez & F. De Brigard (Eds.) *Content and consciousness revisited*. Berlin, GER: Springer.
- Metzinger, T. (2004). *Being no one: The self-model theory of subjectivity*. Cambridge, MA: MIT Press.
- Millikan, R. (1990). The myth of the essential indexical. *Noûs*, 24 (5), 723-734. [10.2307/2215811](https://doi.org/10.2307/2215811)
- (1995). *White queen psychology and other essays for Alice*. Cambridge, MA: MIT Press.
- Moore, G. (1910). The subject-matter of psychology. *Proceedings of the Aristotelian Society*, 10, 36-62.
- Neander, K. (1998). The division of phenomenal labor: A problem for representational theories of consciousness. *Noûs*, 32 (S12), 411-434. [10.1111/0029-4624.32.s12.18](https://doi.org/10.1111/0029-4624.32.s12.18)
- Northoff, G. (2013). Brain and self - a neurophilosophical account. *Child and Adolescent Psychiatry and Mental Health*, 7 (28), 28-28. [10.1186/1753-2000-7-28](https://doi.org/10.1186/1753-2000-7-28)
- Papineau, D. (2002). *Thinking about consciousness*. Oxford, UK: Oxford University Press.
- Pereboom, D. (2011). *Consciousness and the prospects of physicalism*. Oxford, UK: Oxford University Press.
- Prufer, T. (1975). An outline of some Husserlian distinctions and strategies, especially in the Crisis. *Phänomenologische Forschungen*, 1, 189-204.
- Revonsuo, A. (2006). *Inner presence: Consciousness as a biological phenomenon*. Cambridge, MA: MIT Press.
- Rosenthal, D. (1986). Two concepts of consciousness. *Philosophical Studies*, 49 (3), 329-359. [10.1007/BF00355521](https://doi.org/10.1007/BF00355521)
- (1997). A theory of consciousness. In N. Block, O. Flanagan & G. Güzelde (Eds.) *The nature of consciousness: Philosophical debates* (pp. 729-753). Cambridge, MA: MIT Press.
- (2005). *Consciousness and mind*. Oxford, UK: Oxford University Press.
- (2011). Exaggerated reports: Reply to Block. *Analysis*, 71 (3), 431-437. [10.1093/analys/anr039](https://doi.org/10.1093/analys/anr039)
- (2012). Higher-order awareness, misrepresentation and function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367 (1594), 1424-1438. [10.1098/rstb.2011.0353](https://doi.org/10.1098/rstb.2011.0353)
- Rudrauf, D., Bennequin, D., Landini, G. & Williford, K. (ms). Phenomenal consciousness has the form of a projective 3-space under the action of the general projective linear group.
- Sebastian, M. (forthcoming). Experiential awareness: Do you prefer it to me? *Philosophical Topics*.
- Shoemaker, S. (1968). Self-reference and self-awareness. *Journal of Philosophy*, 65 (19), 555-567. [10.2307/2024121](https://doi.org/10.2307/2024121)
- Smullyan, R. (1984). Chameleonic languages. *Synthese*, 60 (2), 201-224. [10.1007/978-94-017-1592-8_11](https://doi.org/10.1007/978-94-017-1592-8_11)
- Stoljar, D. (2006). *Ignorance and imagination: The epistemic origin of the problem of consciousness*. Oxford, UK: Oxford University Press.
- Strawson, G. (2006). Realistic monism: Why physicalism entails panpsychism. In A. Freeman (Ed.) *Consciousness and its place in nature: Does physicalism entail panpsychism?* (pp. 3-31). Charlottesville, VA: Imprint Academic.
- (2009). *Selves: An essay in revisionary metaphysics*. Oxford, UK: Oxford University Press.
- (2011). *The evident connexion: Hume on personal identity*. Oxford, UK: Oxford University Press.
- Thiel, U. (2011). *The early modern subject: Self-consciousness and personal identity from Descartes to Hume*. Oxford, UK: Oxford University Press.
- Tononi, G. (2014). Consciousness: Here, there, but not everywhere. 20th Anniversary Toward a Science of Consciousness April 21-26, 2014 Tucson, Arizona.
- Tononi, G. & Koch, C. (2008). The neural correlates of consciousness. *Annals of the New York Academy of Sciences*, 1124 (1), 239-261. [10.1196/annals.1440.004](https://doi.org/10.1196/annals.1440.004)
- Tye, M. (1995). *Ten problems of consciousness: A representational theory of the phenomenal mind*. Cambridge, MA: MIT Press.
- (2000). *Consciousness, color, and content*. Cambridge, MA: MIT Press.
- (2011). *Consciousness revisited: Materialism without phenomenal concepts*. Cambridge, MA: MIT Press.
- Weisberg, J. (2008). Same old, same old: The same-order representation theory of consciousness and the division of phenomenal labor. *Synthese*, 160 (2), 161-181. [10.1007/s11229-006-9106-0](https://doi.org/10.1007/s11229-006-9106-0)

- (2011a). Misrepresenting consciousness. *Philosophical studies*, 154 (3), 409-433. [10.1007/s11098-010-9567-3](https://doi.org/10.1007/s11098-010-9567-3)
- (2011b). Abusing the notion of what-it's-like-ness: A response to Block. *Analysis*, 71 (3), 438-443. [10.1093/analys/anr040](https://doi.org/10.1093/analys/anr040)
- (2012). On HOTs and HOTIEs: Higher-order thoughts, indexed essentially. 10th Biennial Toward a Science of Consciousness April 9-14, 2012 Tucson, Arizona.
- (2014). *Consciousness*. Cambridge, MA: Polity Press.
- Wilberg, J. (2010). Consciousness and false HOTs. *Philosophical Psychology*, 23 (5), 617-638. [10.1080/09515089.2010.514567](https://doi.org/10.1080/09515089.2010.514567)
- Williams, P. (2000). *The reflexive nature of awareness: A tibetan madhyamaka defence*. Delhi, India: Motilal Banarsidass Publishers.
- Williford, K. (2004). Moore, the diaphanousness of consciousness, and physicalism. *Metaphysica*, 5 (2), 133-153.
- (2005). The intentionality of consciousness and consciousness of intentionality. In G. Forrai & G. Kampis (Eds.) *Intentionality: Past and future* (pp. 143-155). Amsterdam, NL: Rodopi.
- (2006). The self-representational structure of consciousness. In U. Kriegel & K. Williford (Eds.) *Self-representational approaches to consciousness* (pp. 111-142). Cambridge, MA: MIT Press.
- (2007). The logic of phenomenal transparency. *Soochow Journal of Philosophy*, 16, 181-195.
- (2011a). Auto-representacionalismo y los problemas de la subjetividad. *Cuadernos de Epistemología*, 5, 39-51.
- (2011b). Pre-reflective self-consciousness and the autobiographical ego. In J. Webber (Ed.) *Reading Sartre: On phenomenology and existentialism* (pp. 195-210). London, UK: Routledge.
- (2013). Husserl's hyletic data and phenomenal consciousness. *Phenomenology and the Cognitive Sciences*, 12 (3), 501-519. [10.1007/s11097-013-9297-z](https://doi.org/10.1007/s11097-013-9297-z)
- Williford, K., Rudrauf, D. & Landini, G. (2012). The paradoxes of subjectivity and the projective structure of consciousness. In S. Miguens & G. Preyer (Eds.) *Consciousness and subjectivity* (pp. 321-353). Frankfurt, GER: Ontos Verlag.
- Zahavi, D. (1999). *Self-awareness and alterity: A phenomenological investigation*. Evanston, IL: Northwestern University Press.
- (2005). *Subjectivity and selfhood: Investigating the first-person perspective*. Cambridge, MA: MIT Press.
- (2006). Thinking about (self-) consciousness: Phenomenological perspectives. In U. Kriegel & K. Williford (Eds.) *Self-representational approaches to consciousness* (pp. 273-295). Cambridge, MA: MIT Press.

Explaining Subjective Character: Representation, Reflexivity, or Integration?

A Commentary on Kenneth Williford

Tobias Schlicht

While Williford puts forward a self-reflexive account of subjective character, which identifies the subject of experience with episodes (or the stream) of consciousness, an alternative account is defended here that identifies the subject of experience with the whole organism. On this latter approach, a mental representation is conscious if its neural substrate is integrated into the overall neuronal state underlying the conscious state of the organism at that time. This approach avoids an important problem arising for Williford's theory, namely the individuation of episodes. This problem is elaborated in greater detail.

Keywords

Consciousness | Integration | Phenomenal character | Representationalism | Subject of experience | Subjectivity

Commentator

Tobias Schlicht

tobias.schlicht@rub.de

Ruhr-Universität Bochum
Bochum, Germany

Target Author

Kenneth Williford

williford@uta.edu

The University of Texas
Arlington, TX, U.S.A.

Editors

Thomas Metzinger

metzinger@uni-mainz.de

Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt

jennifer.windt@monash.edu

Monash University
Melbourne, Australia

1 Introduction

The starting point for this commentary on Williford's article is the *commitment* to subjective character as a defining feature of consciousness. Subjective character is what *makes* a conscious experience conscious, i.e., what *all* conscious experiences have in common in virtue of which we call them conscious. Kriegel (2009, p. 1) has offered a distinction between the *qualitative character* and the *subjective character* as two

important aspects of any conscious experience. If you have a phenomenally conscious sensation of red, then there is something that it is like for you to have it. On the one hand, having this experience feels like *this* (where *this* quality distinguishes it from feeling a sensation of pain, say). On the other hand, it feels like something *for you* (i.e., it is subjective *in the same sense* that all of your other conscious experiences are sub-

jective). Qualitative character is the *distinguishing* mark of conscious experiences with regard to each other; subjective character is the *common mark* of all my conscious experiences. This latter aspect has also often been referred to as the me-ishness, ipseity or mine-ness of consciousness (Block 1995; Zahavi 1999).

Williford recognizes that, sadly, not all philosophers theorizing about consciousness share this commitment to subjective character, and that some formulations of it in terms of mine-ness are misleading in giving rise to objectionable implications about essential entities (Metzinger 2011). But for the purposes of this commentary we can leave aside such controversies and instead start with a shared commitment to this feature, on which this commentary will exclusively focus.¹ As a constraint on a theory of subjective character, Williford maintains that it has to respect (1) the relational structure of consciousness, and (2) the Humean intuition that one of the relata, the subject, remains somewhat invisible and is at least not constituted by a special (additional) entity. His solution, in short, is to peacefully combine these two intuitions by *identifying* the subject with (an episode of or) the stream of consciousness, which is itself reflexively self-aware. A further claim is that this account is supposedly “compatible with physicalism” (Williford this collection, p. 1). I do not address this aspect of Williford’s rich paper in this commentary, mostly for reasons of space but also because I think that the putative truth of physicalism should not put any a priori constraints on a theory of consciousness.

¹ One might argue that this move is already problematic since it looks like a *petitio principii*. But I simply take it as an analysis of Nagel’s phrase that *there is something it is like for the organism* to experience something red, say. As a characterization of phenomenal consciousness this is almost unanimously accepted in the field. What it picks out according to the present analysis is a variant aspect that differs in different experiences (qualitative character), and an invariant aspect that remains identical across different experiences (subjective character). I do not have enough space here to argue in detail for this analysis. A further reason for distinguishing both aspects, subjective and qualitative character, is the phenomenal observation that we can become *conscious of ourselves as the identical subject* in contrast to the constantly changing stream (or ensemble) of conscious representations. Here, the qualitative differences of the multiple representations we have at a time do not matter. What matters here is that they are related to myself such that I can call them and experience them as mine (cf. Schlicht 2011).

My commentary is thus structured as follows. In the [second](#) section, I will recapitulate Williford’s take on subjective character and point to problems with his identification of the subject with the stream (or episodes) of consciousness. In the [third](#) and [fourth](#) sections, I will present an alternative way of conceptualizing the subject in the context of a theory of consciousness that also satisfies the constraints mentioned above. On this alternative view, a mental representation is conscious (i.e., it exhibits subjective character) if it is integrated in the right way into the overall conscious state of the organism. This overall state includes representations of the state of the organism. By way of integration, all *conscious* representations are something *for* the organism that is identified as the subject of experience. This alternative, which is an instance of an integration-theory, has the advantage both of bypassing the problems that seem to beset Williford’s account and of being not only compatible with but also supported by the best empirical hypotheses about consciousness currently available. I will sketch an argument for this view and attempt to answer possible objections to the premises of this argument.

2 Williford on subjective character

Williford’s aim is to characterize the subjective character of consciousness in a way that accounts “for both the Humean intuition that the subject-relatum is, in some sense, invisible and that, nevertheless, consciousness has a subject-object relational structure that is phenomenally manifest and non-inferentially knowable” ([this collection](#), pp. 10-11). There are three constraints on an account of subjective character, according to Williford: (a) conscious experiences are relational in having both a subject- and an object-pole; (b) the subject-pole is not constituted by some additional, irreducible, or otherwise special entity; (c) the subject-pole must be something that is nevertheless manifest in consciousness, not hidden from it.

Note that it is not an option for someone taking subjective character seriously to agree that phenomenal consciousness is relational, in-

volving a subject-pole, but at the same time holding that this pole is forever “*hidden*” (p. 11). This does not work because such an account could not explain subjective character. After all, we consider subjective character as real *only* because it supposedly shows up phenomenally: I experience my conscious states as *mine*.²

Williford’s commitment to a subject (subject-pole) rules out the possibility that experiences may be free-floating entities, not being enjoyed by *anyone*. Phenomenal consciousness is supposed to be relational through-and-through, directed *at* some object and existing *for* some subject: “anything that phenomenally appears, appears to someone or something” (p. 9).³ In general, Williford attempts to capture both the intentionality and the subjectivity of consciousness in the slogan that every experience involves the “appearance of something to something” (p. 9), where the latter refers to subjectivity. He leaves the notion of “*appearing* or of *phenomenally manifesting* undefined” (p. 10), but in order for what he says to make sense we have to take it to be just another way of saying that something is *phenomenally conscious*: it is “just the appearance to/in consciousness of something” (p. 10).

In order to meet the constraints he set for himself, Williford identifies the subject with the stream of consciousness or with (some complex or rich) episode of consciousness (p. 10). This identity claim then leads to the situation that the subject-pole of the consciousness-relation *appearing* (or being manifest) in the conscious episode is the episode itself. The subject-pole is thereby manifest, i.e., consciously experienced, but not separable as an entity from the conscious episode in question, and thus it is—in a sense—invisible. But it is only invisible in the sense that there is no additional entity that ac-

counts for the subject-pole. In order to meet the constraints mentioned above, Williford therefore defends “the view that consciousness is self-manifesting” (p. 10), i.e., an episode or stream of consciousness *appears to itself* no matter what else is manifest to consciousness (some perceived object, say).

Partly because Williford subscribes to the Humean intuition that we do not find a self, or a “self-entity, me-haecceity, me-ish quale, or subject-relatum” (p. 10) if we turn to our stream of conscious experiences, he is led to the identification of the subject-relatum with the stream of consciousness itself. Although the conscious episode appears itself *in* the episode, consciousness is self-reflexive, yet not self-representing. The relevant difference between an unconscious and a conscious episode is not due to some form of representation. Rather, the conscious episode contains an internal relational (intrinsic) property that is responsible for the episode’s being *acquainted with itself*.⁴ Subjective character is thus supposedly “the self-acquaintance of every instance of consciousness” (p. 1), which these instances exhibit in virtue of “some internal relational property” (p. 1). The subject of experience, being *identical* to the episode of consciousness, is self-acquainted. But although consciousness is self-reflexive, the claim is not that a mental episode becomes conscious through an act of reflection directed at it (p. 10). This is an impossible path when it comes to explaining subjective character, since an act of reflection presupposes that what it reflects upon is already *mine* in the relevant sense to be explained (Frank 2007; Zahavi 1999). Reflection can discover but not bring into being a self-referential conscious state.

Now, the stream (or episode) of consciousness exhibits subjective character in the sense that the stream itself is manifest within the stream so that the relationality constraint is met, although no additional entity need be introduced in order to play the subject-role. Therefore, the Humean invisibility-constraint is met as well. This is more or less the positive

² One way to put this with respect to sensations like hunger is to say that, since they are related to me in such an unmediated sense, it is impossible to be mistaken about the subject undergoing such sensations (Shoemaker 1968).

³ This claim is defended especially in opposition to what Williford calls F-theories, or varieties of first-order representationalism such as Tye’s (1995) PANIC-theory, which arguably neither accepts nor explains subjectivity so understood. Higher-order and same-order accounts at least accept this feature of consciousness, which they—mistakenly—attempt to explain in terms of representation.

⁴ Thus, the property of being conscious (and thus subjective) is not bestowed upon the episode by some external property, like a higher-order thought directed at (or representing) it (Rosenthal 2005).

story as far as I have understood it. The main philosophical problem for Williford's account is to formulate criteria as to how to individuate an episode. This problem leads to a dilemma for his account that is spelled out in more detail below.

If we follow Williford and identify the subject with complex conscious episodes (or even *the whole* stream of consciousness), then subjective character only seems to arise for complex episodes, and not for any of the episode's parts or elements: "[t]he episode is a unified whole, the differentiated qualities and objects appearing in/to it are like its parts [...]" (pp. 10-11). Since he emphasizes that all episodes have parts (*ibid.*), I take it that a single sensation of red, say, consequently does *not* count as an episode, because it can hardly be separated into parts; then it can instead always appear only as an element of an episode which is in turn a "unified whole". On the other hand, Williford also emphasizes that, trivially, everything always also is an improper part of itself. On this reading, a single sensation of red *could* be an episode. This gives rise to the following options regarding the individuation of episodes that can be put in terms of a dilemma:

1. *If a single sensation of red is too simple to count as an episode*, then all that Williford's theory can explain is why the complex episode as an emergent whole (having single experiences as its parts or elements) is conscious. It *cannot* explain what makes an individual element of this whole episode (or stream), a sensation of red say, conscious. But the varieties of representationalism (which he criticizes) aim to explain exactly this feature of consciousness. A problem with this first horn of the dilemma is thus that we need to answer the question whether or not such single sensations can be conscious independently of being an element of a larger episode.
 - a) *If individual sensations can be conscious independently*, then the question arises as to whether they can be conscious *without* thereby exhibiting subjective character (given subjective character only arises on

the level of whole episodes). This is not what Williford should accept since he takes subjective character to be a defining feature of consciousness; there is no consciousness without subjective character. So if an individual sensation of red could be conscious then it could be so only by exhibiting subjective character. This seems to lead us to Zeki's theory of "micro-consciousness" (Zeki & Bartels 1998; Zeki 2007) according to which every individual node of a perceptual system (visual, auditory etc.) can generate an "atom" of consciousness independently. This is an extreme version of what Bayne (2010) calls an "atomistic" approach to consciousness, standing in contrast to more "holistic" approaches:

"Theorists that adopt an atomistic orientation assume that the phenomenal field is composed of 'atoms of consciousness'—states that are independently conscious. Holists, by contrast, hold that the components of the phenomenal field are conscious only as the components of that field. Holists deny that there are any independent conscious states that need to be bound together to form a phenomenal field. Holists can allow that the phenomenal field can be formally decomposed into discrete experiences, but they will deny that these elements are independent atoms or units of consciousness." (Bayne 2010, pp. 225-226)

The problem with such atomistic approaches is really the phenomenon of the unity of consciousness, i.e., that such individually conscious units would need to be bound together to a much larger all-encompassing unified "phenomenal field", as Bayne puts it, in order to account for what we actually experience. But then we should expect there to be a *mechanism* responsible for such phenomenal binding, a mechanism that we also should expect to break down occasionally under certain circumstances; but there is no evidence for

such a mechanism. The phenomenal unity of consciousness seems to be a deep feature of consciousness just like subjective character, in the sense that it cannot break down and that phenomenal consciousness cannot occur without it. I agree with Bayne's point here (cf. [Schlicht 2007](#)), and I think that Williford would not be prepared to take Zeki's route either. At least there is no indication in the text that would support this reading. Alas, Williford also sets aside the important issue of the unity of consciousness, which arises given the unresolved problem of providing criteria for the individuation of episodes.

- b) So we are left with the alternative that *individual sensations cannot be conscious independently*. For an individual element to become conscious (and to exhibit subjective character) it must then be *integrated* into a larger (cumulative) episode. What's needed then is a theory (and a mechanism) explaining how such integration into an episode takes place. However, then we are left with an alternative view regarding the question of what is responsible for a representation's being conscious, namely some kind of integration-theory. In fact, that is the path I will recommend (and elaborate in more detail) below in section 3. The general idea is that phenomenally-conscious representations are those that are adequately integrated into a global state (we may call it an episode). My worry with regard to Williford's account is simply that once we have such an integration-account, there is no need for his additional story in terms of self-reflexivity in order to explain subjective character. Since subjective character is (taken to be) a defining feature of conscious experience, an account that informs us about how individual sensations become conscious will also inform us about how they acquire subjective character: through integration.

2. But that's not the end of the story. Williford simply could say that a sensation of red may

be a conscious episode. So far, we have discussed the problem of individuating episodes on the assumption that a single sensation of red cannot count as an episode. Now we have to discuss the consequences of the assumption that *a single sensation of red may count as an episode*. This leads to two further possibilities.

- a) One could accept such minimal episodes despite the fact that this concession gives rise to a multiplicity of (streams and consequently) conscious subjects. Although it's metaphysically (somewhat) extravagant, this is a perfectly coherent position to take. Indeed, it seems to be akin to Strawson's theory of the self, according to which a self lasts only as long as an individual state (or episode) of consciousness ([Strawson 1997](#)). But this view flies in the face of experience. For one thing, it is inadequate to explain an important aspect of consciousness, namely what we may call, following Kant, the *(empirical) consciousness of the identity of oneself as subject*: "I am [...] conscious of the identical self in regard to the manifold of the representations that are given to me in an intuition because I call them all together *my* representations, which constitute *one*" ([B134](#)). What he means is that, at least in non-pathological cases, I can become conscious of myself *as* the single, (synchronically as well as diachronically) identical subject vis-à-vis my diverse experiences. I *never* identify myself with one or many of my conscious representations (or episodes for that matter). Rather, I distinguish myself from them *as the subject* who has them when I self-ascribe them. And this empirical consciousness of an identical subject is possible, according to Kant, *because* all my *conscious* experiences are already self-related. I can already call them *mine* because they exhibit subjective character simply by being phenomenally conscious. Kant, famously and notoriously, tried to account for this consciousness of self by simply postulating a *transcendental unity* of apperception in which this is sup-

posed to originate. If Strawson's view were correct, then Kant would presumably reply by pointing to a natural, yet implausible consequence: "I would have as multicoloured, diverse a self as I have representations of which I am conscious" (CpR B134). Accepting this horn of the dilemma therefore has the consequence that we would now need a story that helps us make sense of how the subject of the sensation of red is related to the subject that is identified with an auditory sensation of a loud sound, etc. In effect, this would lead to a binding problem for the multitudinous "subjects" of experience, since in my view, we cannot be content with a multiplicity of conscious subjects. I also think that Williford might not be satisfied with such an outcome, since he never entertains the possibility of multiple subjects in his essay.

- b) Therefore—again, on the hypothesis that a single sensation of red counts as an episode—one could argue that *the multiplicity of conscious episodes has to be overcome in favor of one (unified) stream of consciousness*. This calls, again, for an integration mechanism that produces such a unity. Though I can understand why one would now identify this resulting integrated single stream of consciousness with the subject of experience, I don't see any motivation to identify the episode "single sensation of red" with a subject of experience, if a more complex combination of episodes is needed anyway.

I conclude that the problem of individuating episodes either leads to the acceptance of implausible views like Zeki or Strawson's theories of consciousness and self or to the need for an integration account that explains how individual elements are combined into the one global conscious experience. The claim I would like to put forward is that once we have such an integration account, Williford's proposal becomes superfluous, because what it is intended to explain is then already explained by the integration account.

3 Integration vs. representation

When the aim is to provide an account of the difference between a representation's being phenomenally conscious and it's being unconscious many philosophers are drawn to some form of *representationalism*. This is motivated in part by the prospect of reducing the problem of consciousness to the problem of intentionality or *representation* (Tye 1995; Dretske 1995; Rosenthal 2005; Lycan 1996; Metzinger 2003; Kriegel 2009; Kriegel & Williford 2006). But many of those who are dissatisfied with a representational criterion argue that the difference is due to some sort of *integration* (Dehaene 2014; Van Gulick 2004; Edelman & Tononi 2000; Damasio 2010; Metzinger 1995; Kant 1999; Schlicht 2011). Such integration may eventually result in a higher-order or more complex representational state. In that sense, the two accounts do not mutually exclude each other. But they give different answers to the question of what is responsible for the representation exhibiting the feature of being conscious. To put forward both a representational condition and an integration mechanism would amount to wearing a belt as well as suspenders. Williford's paper demonstrates that other theories are also possible. He favors self-reflexivity as the core feature a representation must exhibit in order for it to be conscious.

In the first part of his paper, Williford scrutinizes all dominant varieties of representationalism, especially with respect to their explanatory power regarding the subjective character of conscious experiences. His case against first-order, higher-order, and same-order or self-representationalism is solid, and I have nothing to add in this regard (cf. also Schlicht 2008b; Vosgerau et al. 2008).⁵

The basis for answering the question as to which conditions have to be met by a single sensation of red in order for it to be conscious and subjectively experienced is the observation

⁵ I disagree with respect to what Williford calls P-Theories, according to which a "privileged object" is represented which makes all the difference between conscious and unconscious representations. Williford interprets Damasio's theory in this way, but although various representations (of the body especially) play an important role in Damasio's theory (as in most other theories), this is not the whole story (see fn. 7).

that the organism in question is already conscious in the creature-sense. This general consciousness (or state of vigilance) admits of degrees (from deep coma to wakefulness) and is one of the conditions for being able to enjoy a sensation of red at all (Dehaene et al. 2006). Empirical evidence points to the assumption that the neural structures in the brain supporting this state contain the relevant structures monitoring and regulating the homeostatic balance of the whole organism. Damasio (1999, 2010) calls these structures “proto-self”-structures, the biological forerunner of that which we eventually experience as a sense of self. He assumes that the brain can only perform these functions of monitoring and regulating if the overall state of the whole organism is represented in the brain.

In addition to representations of the organism, the brain is assumed to produce representations of (objects in) the external world. Given the limited capacity of conscious perception and memory systems, such representations stand in competition (Koch 2004). The basic idea of integration-theories is that some of these competing representations, like a sensation of red, are conscious because they are integrated into a more global state that also contains the structures responsible for creature-consciousness. Van Gulick (2004) has sketched such an integration-theory, based on ideas already to be found in Metzinger (1995):

The basic idea is that lower-order object states become conscious by being incorporated as components into the higher-order global states (HOGS) that are the neural and functional substrates of conscious self-awareness. The transformation from unconscious to conscious state is not a matter of merely directing a separate and distinct meta-state onto the lower-order state but of “recruiting” it into the globally integrated state that is the momentary realization of the agent’s shifting transient conscious awareness. (Van Gulick 2004, pp. 76-77)

In other words, a single sensation of red is consciously experienced if the neural activation

pattern supporting this sensation is integrated in the right way into the neural basis representing the overall state of the organism, the “dynamic core” in Edelman’s words (Edelman & Tononi 2000).⁶

Importantly, the integration mechanism (which is what has to be determined empirically in this framework)—synchronous oscillations, say—is not only responsible for producing a coherent single experiential state of the organism; it also thereby conveys subjective character to the integrated individual representations. If this idea is combined with Damasio’s (1999) notion of proto-self-structures, then integration facilitates a strong connection between the substrate of an individual sensation (of red, say) and the biological structure representing the organism in the brain.⁷ Of course, just like on all other theories, the hard problem is not addressed head-on, i.e., it is not explained *why* activation of these structures feels like something at all. All that can be provided (at this stage anyway) is a coherent story of how all these aspects hang together. But one advantage of the present integration-account is that by establishing a connection between the organism (as represented in the brain) and its object-representations we can make sense of the important fact that all conscious representations feel like something *for the organism*. The organism provides, as Damasio puts it, a “haven of stability and invariance” (1999, p. 142, p. 153; see also Metzinger 2003, p. 161), i.e., just what we need in order to account for subjective character. For remember that subjective

6 Another way to think of this is along the lines of the “Global Neuronal Workspace Model” in which attentional mechanisms determine which of the neural coalitions are integrated (Dehaene et al. 2006). But means other than attention are possible.

7 Williford discusses Damasio’s theory under the label of a P-Theory as a variety of representationalism and finds it wanting. Of course, representations of various sorts, especially of the organism, play an important role in Damasio’s theory (as in many other theories). But I do not share Williford’s interpretation that it is these (special) representations as such that are responsible for consciousness. Various representations (or maps, as Damasio also calls them) have to be integrated in the right kind of way in order for there to be something it is like for the organism. Therefore, I do not consider Damasio’s theory a version of representationalism since there, the mechanism responsible for consciousness is not representation but integration of body representations with object representations via recurrent activations in so-called “convergence zones” (Damasio 1994, p. 95-96, 162).

character is the feature that remains stable across different representations, while qualitative character is the feature that distinguishes different representations from each other. So in order to get an account of subjective character started, we have to look for the point of “*maximal invariance of content*” in the conscious model of reality”, as Metzinger (2003, p. 134) puts it. Metzinger agrees that this invariance is most likely due to the organism and its bodily structures represented in the brain, since it is invariance (or maintenance of homeostatic balance) that keeps the organism alive. Another advantage of this view is that it does so without introducing a questionable new entity and by avoiding Williford’s phenomenologically counterintuitive claim that the stream of consciousness should be identified with the subject of experience. In this commentary, I cannot argue in detail for this positive alternative but hope that these sketchy comments give the reader a general idea of what it aims at. Since I am dissatisfied with Williford’s identification of the subject of experience with the stream or an episode of consciousness, let me now finally turn to an argument for a different conceptualization of the subject.

4 The subject as organism

My alternative claim is that we should simply identify the subject with the organism. This section is an attempt to support this bold claim. The premises of the argument focus on analyses of the structures of phenomenal consciousness and intentionality:

Premise 1 (phenomenal consciousness):

Phenomenal consciousness is characterized by there being something that it is like *for a subject* to be in that state. In this minimal sense, consciousness is relational and requires the assumption of a *subject-pole* of experience.

Premise 2 (intentionality):

The structure of intentionality is such that a *subject* is directed (via some psychological act or attitude like believing, desiring, perceiving etc.) at a content, object, or state of affairs. Inten-

tionality is quasi-relational since at least the *subject* must exist, although the intentional object need not exist.⁸

Premise 3 (subject identity):

The subject that is intentionally directed is *identical* to the subject for whom there is something that it is like to be in a given mental state.

Premise 4 (embodied cognition):

Many intentional attitudes (like perceiving, grasping, emoting) are *embodied* and can be ascribed only to an embodied agent, i.e., to the whole organism.

Conclusion:

The *subject* for which there is something it is like to be in a given mental state and the *subject* that is intentionally directed at a content or object is the *organism*.

4.1 Elaboration of the premises

Premise 1: Phenomenal consciousness

First of all, it is interesting to note that Nagel’s initial characterization of consciousness in terms of *there being something that it is like* is already concerned with *the organism* as the entity *for which* there is something that it is like: After having noted the diversity of beings capable of conscious experience which may lead to very different kinds of conscious experience, Nagel argues that “no matter how the form may vary, the fact that an organism has conscious experience *at all* means, basically, that there is something it is like to *be* that organism, [...] something it is like *for* the organism” (1974, p. 436). So, given that the philosophical community seems to have agreed to refer to Nagel’s slogan in order to characterize phenomenal consciousness in the first place, they should seriously consider Nagel’s talk of the organism as the subject of experience. But apart from this ob-

⁸ A reviewer pointed out that different thinkers had different opinions about *what* is intentionally directed: the mental state, the psychological act, or the thinker etc. As will become clear below, I do not share the view that a mental state is itself directed, but favor the view that a creature of some sort is directed at something *via* an act or attitude. A great advantage of this view is that such attitudes are not limited to mental states like beliefs and desires (as traditionally held), but it also allows also for motor intentional attitudes like grasping or holding etc., i.e., essentially bodily ways of being directed (premise 4). For details see Schlicht (2008a).

servation, all that is stressed in the first premise is the relational character of phenomenal consciousness, much in the sense of one of the commitments defended in Williford's paper. The reasons for holding this are mainly phenomenological: it simply appears that way. And we are all aiming at a theory of why this is so. Williford's elaboration of the relational structure of consciousness in terms of the genitive and dative of manifestation captures the intuition expressed in this premise very well. Thus, there is not much room for disagreement here.⁹

Premise 2: Intentionality

In his canonical elaboration of the structure of intentionality, Subject—Intentional Mode—Content, [Tim Crane \(2001, p. 31\)](#) admits that he does not provide an account of the first relatum, "because the nature of the subject is not something that is within the scope of this book (strange as that may seem)". Yet, as far as intentional states are concerned, the assumption that attitudes are not free-floating entities but come along with a thinker, perceiver, or believer is rather uncontroversial. What's controversial is how we should characterize the subject and what kind of commitment is implied in the "acceptance" of a thinker, perceiver, or believer.¹⁰

Premise 3: Subject identity

In a way, this premise is at the same time trivial and important. First of all, if one accepts premises [1](#) and [2](#), then it is natural to accept premise [3](#), if only because the alternative would lead to a multiplicity of subjects, giving rise to questions regarding the relations between them. I discussed this option above in section [2](#). There are many debates about the relation between consciousness and intentionality, but there is hardly any debate about the relation between the subjects of each. So in a way, this premise simply states the obvious, given premises [1](#) and [2](#). But it is plausible to accept it even independently of these premises as the default position. One important reason for this is that there are many conscious experiences that are both phenomenal and intentional—perceptual experi-

ences, for example. If I am looking at a red tomato, then my conscious experience presents me with an object in the external world at which I am thereby visually directed. But there is also something that it is like for me to see the tomato if I am phenomenally conscious of it. Since it would be odd to claim that there are two subjects involved here—one being intentionally directed and one being conscious of the tomato—the default position is that it is one the same subject that is intentionally directed and phenomenally conscious. Second, despite the discussion among analytic philosophers in the last fifty years, it is not clear that phenomenal consciousness and intentionality can be separated from each other so easily anyway. In fact, proponents of phenomenal intentionality (or cognitive phenomenology, see [Bayne & Montague 2011](#)) like [Searle \(1992\)](#), [Strawson \(2004\)](#), [Pitt \(2004\)](#), [Horgan & Tienson \(2002\)](#), [Kriegel \(2013\)](#) and others argue to the contrary. Again, then the premise simply states the obvious.

But this premise also is important because once we commit to it, we can follow either premise [1](#) or [2](#) in our investigation to see whether we can formulate constraints on the nature of the subject based on either consciousness or intentionality. This is the job of premise [4](#), which accepts lessons from recent investigations into ways of being intentionally directed.

Premise 4: Embodied Cognition

Cognitive Science has recently been dominated by discussions on the so-called 4Es, i.e., embodied, embedded, enactive, and extended cognition. These notions characterize four important ways in which our current theorizing about cognition departs from classical cognitive science. They are more or less independent of each other and can be accepted and rejected in isolation.¹¹ This is not the place to elaborate in detail all four of them, especially because for the purposes of this argument only the feature of embodiment is important. Many of our psychological acts, like *perceiving*, being *emotionally* directed at or affected by something or other, performing *intentional actions*, etc., are embodied

⁹ I support this premise in more detail in [Schlicht \(forthcoming\)](#). For the purposes of this commentary it is sufficient to note the agreement on the intuition that phenomenal consciousness is relational.

¹⁰ Again, I argue for this premise in [Schlicht \(forthcoming\)](#).

¹¹ An exception may be the intricate connection between cognition being embodied and (therefore) being embedded.

in the sense that features of an organism's non-neural body contribute importantly—be it causally or even constitutively—to the execution of these cognitive acts (Wilson & Foglia 2011).

A plausible claim defended by enactivists is that even a basic cognitive act like perceiving involves many bodily movements like eye-, head- and whole-body movements when looking at or focusing on an object, or when jointly attending to an object with someone else (Noë 2004). This can be accepted independently of more radical claims regarding the usefulness of representations typically put forward by enactivists (Hutto & Myin 2013). What's more, a bulk of empirical evidence has accumulated that supports the important role of the body and bodily actions for psychological acts:

- a) Facial expressions and bodily postures are arguably constitutive elements of feelings and their expression. Many theories of emotion such as multifactorial models (e.g., Scherer 2009; Welpinghus 2013) usually include as one component a bodily feature. Moreover, eye- and head-movements count among the constitutive and content-determining elements of visual perception (Noë 2004).
- b) Research on mirror neurons has demonstrated the intricate relation between perceiving and acting in the sense that the same neural structures are employed for the execution and observation or recognition of intentional acts and emotional expressions (Rizzolatti & Sinigaglia 2008; Keysers 2013). Controversial debates about the role of mirror neurons for social cognition notwithstanding, it is fair to say that from a neural perspective, perception and action have to be considered as constituting one single complex system. We develop motor programs for the performance of certain actions and *reuse* these programs in our observation of others when they perform such actions. These motor programs contain goal-directed representations with a bodily format (Goldman & de Vignemont 2009) that are crucially different from the propositional format of a belief, say.
- c) What's more, lessons from studies of pathological conditions like visual form agnosia (Mil-

ner & Goodale 1995) suggest that we can be directed at an object in a purely motor-intentional way, thereby demonstrating a “bodily understanding” (Kelly 2002) of an object that is not based on concepts and cannot be put into appropriate words.

Generalizing these (and many other) points (see e.g., Gallagher 2005) leads to a paradigm shift with regard to our understanding of the subject of intentionality: intentionality is not restricted to propositional attitudes; an embodied agent, i.e., an organism,¹² has many sensorimotor, affective, and cognitive means to be directed at objects and states of affairs. This way of understanding the structure of intentionality allows us to capture many more phenomena that clearly fall under the name of intentionality as directedness, e.g., reaching for and grasping an object.

All the premises taken together yield the conclusion that there is one subject capable of intentionality and consciousness that can be identified with the organism (not with the stream of consciousness), characterized by a variety of cognitive capacities allowing for a range of intentional attitudes—some of which are affective,¹³ others sensorimotor,¹⁴ and still others are of sophisticated cognitive¹⁵ varieties. The overall state of the subject, being the whole organism, is represented in the brain. This representation contains information about its body, its interior milieu, etc., such that all representations having to do with the organism's interaction with objects can be coupled to or integrated with the representations monitoring and regulating the state of the organism in the

¹² Talk about embodied agents is broader than talk about organisms. The biological constraints on full-blown cognitive and conscious agents are currently unknown. Whether artificial cognitive systems are possible depends on the limits set by such constraints. In this paper, I cannot address this point.

¹³ One of Brentano's examples in his famous passage on intentionality being the mark of the mental is love, in which someone is loved. This example cannot be adequately captured by restricting intentionality to propositional attitudes which can be formulated using “that-clauses”.

¹⁴ Many forms of being intentionally directed are sensorimotor, e.g., all that has to do with perception and action, this being the biologically primary form of intentionality (Searle 1983, p. 36).

¹⁵ Most cognitive varieties of intentionality are sophisticated and propositional, like beliefs and desires, which can be put into sentences containing “that-clauses”, e.g., Ken believes that physicalism is true.

brain. On the basis of such couplings, it is in principle possible to make sense of the idea that object-representations become subjective in the sense of being something *for* the organism.

A caveat: this does not amount to an explanation of how consciousness arises in the first place, or of *why* integrated representations are experienced at all. But hardly any theory of consciousness has properly addressed this hard problem so far (Chalmers 1996). The limited claim of this commentary is that we can at best make sense of the subjective character of phenomenal consciousness if we adopt an integration-theory as outlined above and regard the subject for which there is *something that it is like* as the whole organism. As I concluded in the first part, depending on how he is going to individuate episodes—a problem which he has not yet solved—, Williford seems to be in need of such an integration-account anyway. Therefore, this sketch of an alternative should be appealing for someone taking subjective character seriously.¹⁶

Acknowledgements

I would like to thank Jennifer Windt, Thomas Metzinger, and two anonymous reviewers for very helpful comments on an earlier draft of this paper. I am also grateful to Kenneth Williford, who provided such a stimulating and challenging paper and once again to Jennifer Windt and Thomas Metzinger for the chance to contribute this commentary.

This work is funded by the Volkswagen Foundation.

References

- Bayne, T. (2010). *The unity of consciousness*. Oxford, UK: Oxford University Press.
- Bayne, T. & Montague, M. (2011). *Cognitive phenomenology*. Oxford, UK: Oxford University Press.
- Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18 (2), 227-287. [10.1017/S0140525X00038188](https://doi.org/10.1017/S0140525X00038188)
- Chalmers, D. J. (1996). *The conscious mind*. Oxford, UK: Oxford University Press.
- Crane, T. (2001). *Elements of mind*. Oxford, UK: Oxford University Press.
- Damasio, A. (1994). *Descartes' error*. New York, NY: Avon Books.
- (1999). *The feeling of what happens*. New York, NY: Basic Books.
- (2010). *Self comes to mind*. New York, NY: Basic Books.
- Dehaene, S. (2014). *Consciousness and the brain. Deciphering how the brain codes our thoughts*. New York, NY: Viking.
- Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J. & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: A testable taxonomy. *Trends in Cognitive Sciences*, 10 (5), 1-8. [10.1016/j.tics.2006.03.007](https://doi.org/10.1016/j.tics.2006.03.007)
- Dretske, F. (1995). *Naturalizing the mind*. Cambridge, MA: MIT Press.
- Edelman, G. M. & Tononi, G. (2000). *A universe of consciousness: How matter becomes imagination*. New York, NY: Basic Books.
- Frank, M. (2007). Non-objectal subjectivity. *Journal of Consciousness Studies*, 14 (5-6), 152-173. [10.1007/s11098-011-9837-8](https://doi.org/10.1007/s11098-011-9837-8)
- Gallagher, S. (2005). *How the body shapes the mind*. Oxford, UK: Oxford University Press.
- Goldman, A. & de Vignemont, F. (2009). Is social cognition embodied? *Trends in Cognitive Sciences*, 13 (4), 154-159. [10.1016/j.tics.2009.01.007](https://doi.org/10.1016/j.tics.2009.01.007)
- Horgan, T. & Tienson, J. (2002). The intentionality of phenomenology and the phenomenology of intentionality. In D. J. Chalmers (Ed.) *Philosophy of mind. Classical and contemporary readings* (pp. 520-533). Oxford, UK: Oxford University Press.
- Hutto, D. D. & Myin, E. (2013). *Radicalizing enactivism. Basic minds without content*. Cambridge, MA: MIT Press.
- Kant, I. (1999). Critique of pure reason. In A. Wood & P. Guyer (Eds.) Cambridge, UK: Cambridge University Press.

¹⁶ The limits of this commentary do not permit an exhaustive discussion of possible objections to this account, but I discuss it at greater length in Schlicht (forthcoming).

- Kelly, S. (2002). Merleau-Ponty on the body: The logic of motor intentionality. *Ratio*, 15 (4), 376-391.
[10.1111/1467-9329.00198](#)
- Keysers, C. (2013). *The empathic brain*. Groningen, NL: Social Brain Press.
- Koch, C. (2004). *The quest for consciousness*. Englewood, CL: Roberts & Company.
- Kriegel, U. (2009). *Subjective consciousness*. Oxford, UK: Oxford University Press.
- (2013). *Phenomenal intentionality*. Oxford, UK: Oxford University Press.
- Kriegel, U. & Williford, K. (Eds.) (2006). *Self-representational approaches to consciousness*. Cambridge, MA: MIT Press.
- Lycan, W. (1996). *Consciousness and experience*. Cambridge, MA: MIT Press.
- Metzinger, T. (1995). Faster than thought. Holism, homogeneity and temporal coding. In T. Metzinger (Ed.) *Conscious experience* (pp. 425-461). Thorverton, UK: Imprint Academic.
- (2003). *Being no one*. Cambridge, MA: MIT Press.
- (2011). The no-self alternative. In S. Gallagher (Ed.) *The oxford handbook of the self* (pp. 279-296). Oxford, UK: Oxford University Press.
- Milner, A. D. & Goodale, M. A. (1995). *The visual brain in action*. Oxford, UK: Oxford University Press.
- Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, 83 (4), 435-450.
- Noë, A. (2004). *Action in perception*. Cambridge, MA: MIT Press.
- Pitt, D. (2004). The phenomenology of cognition or what is it like to think that p? *Philosophy and Phenomenological Research*, 69 (1), 1-36.
[10.1111/j.1933-1592.2004.tb00382.x](#)
- Rizzolatti, G. & Sinigaglia, C. (2008). *Mirrors in the brain*. Oxford, UK: Oxford University Press.
- Rosenthal, D. (2005). *Consciousness and mind*. Oxford, UK: Oxford University Press.
- Scherer, Klaus R. (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition & Emotion*, 23 (7), 1307-1351.
[10.1080/02699930902928969](#)
- Schlicht, T. (2007). *Erkenntnistheoretischer Dualismus. Das Problem der Erklärungslücke in Geist-Gehirn-Theorien*. Paderborn, GER: Mentis.
- (2008a). Ein Stufenmodell der Intentionalität. In P. Spät (Ed.) *Zur Zukunft der Philosophie des Geistes* (pp. 59-91). Paderborn, GER: Mentis.
- (2008b). Selbstgefühl. Damasio's Stufentheorie des Bewusstseins und der Emotion. In E. Düsing (Ed.) *Geist und Psyche* (pp. 337-369). Würzburg, GER: Königshausen & Neumann.
- (2011). Non-conceptual content and the subjectivity of consciousness. *International Journal of Philosophical Studies*, 19 (3), 489-518. [10.1080/09672559.2011.595197](#)
- (forthcoming). Selves, or something near enough. In S. Grätzel (Ed.) *Buddhism and Bioethics*.
- Searle, J. R. (1983). *Intentionality*. Cambridge, MA: MIT Press.
- (1992). *The rediscovery of the mind*. Cambridge, MA: MIT Press.
- Shoemaker, S. (1968). Self-reference and self-awareness. *Journal of Philosophy*, 65 (19), 555-567.
[10.2307/2024121](#)
- Strawson, G. (1997). The self. *Journal of Consciousness Studies*, 4 (5-6), 405-428.
- (2004). Real intentionality. *Phenomenology and the Cognitive Sciences*, 3 (3), 287-313.
[10.1023/B:PHEN.0000049306.63185.0f](#)
- Tye, M. (1995). *Ten problems of consciousness*. Cambridge, MA: MIT Press.
- Van Gulick, R. (2004). Higher-order global states. In R. Gennaro (Ed.) *Higher-order theories of consciousness* (pp. 67-92). Amsterdam, NL: John Benjamins.
- Vosgerau, G., Schlicht, T. & Newen, A. (2008). Orthogonality of phenomenality and content. *American Philosophical Quarterly*, 45 (4), 329-348.
- Welpinghus, A. (2013). *Emotions as natural and social kinds*. (Unpublished)
- Williford, K. (2015). Representationalisms, subjective character, and self-acquaintance. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.
- Wilson, R. & Foglia, L. (2011). Embodied cognition. *The Stanford Encyclopedia of Philosophy*. E. N. Zalta (Ed.) <http://plato.stanford.edu/entries/embodied-cognition/>
- Zahavi, D. (1999). *Self-awareness and alterity. A phenomenological investigation*. Englewood, IL: Northwestern University Press.
- Zeki, S. (2007). A theory of micro-consciousness. In M. Velmans & S. Schneider (Eds.) *The Blackwell companion to consciousness* (pp. 580-588). Oxford, UK: Blackwell.
- Zeki, S. & Bartels, A. (1998). The autonomy of the visual systems and the modularity of conscious vision. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 353 (1377), 1911-1914.
[10.1098/rstb.1998.0343](#)

Individuation, Integration, and the Phenomenological Subject

A Reply to Tobias Schlicht

Kenneth Williford

Tobias Schlicht argues that subjective character derives from the integration of mental states into a complex of representations of the organism and that therefore there is no need try to account for subjective character in terms of “reflexivity” or self-acquaintance, as I do. He further argues that the proper subject of consciousness is the whole organism and not the episode or stream of consciousness, as I maintain. He maintains that his account solves problems about the individuation and synchronic unity of conscious mental states that mine does not. While I agree that we need an account of the individuation of episodes of consciousness and an account of the synchronic and diachronic unities of consciousness (something I bracketed in my paper), I disagree that making the organism into the phenomenological subject of consciousness helps with these problems. However, I am willing to concede that the organism is the subject of consciousness in some non-phenomenological sense.

Keywords

Conscious vs. unconscious mental states | Individuation | Integration | Organism | Phenomenological subject | Reflexivity | Self-acquaintance | Unity of consciousness

Author

[Kenneth Williford](#)
williford@uta.edu
The University of Texas
Arlington, TX, U.S.A.

Commentator

[Tobias Schlicht](#)
tobias.schlicht@rub.de
Ruhr-Universität
Bochum

Editors

[Thomas Metzinger](#)
metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

[Jennifer M. Windt](#)
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Introduction

In his insightful commentary on my contribution to *Open MIND*, Tobias Schlicht argues for the following claims: (1) The subject of conscious episodes should be identified with the organism whose episodes they are ([Schlicht this collection](#), pp. 2, 8-9). (2) Once we understand how non-conscious mental states (perceptions, thoughts, etc.) become conscious by being integrated into the underlying organismal creature-consciousness, we will have understood all that is important about how a

conscious state is endowed with subjective character ([Schlicht this collection](#), pp. 5-6). And, (3), such an account would obviate an account like mine, since there would be no need to imagine that individual episodes of consciousness have a sort of self-contained subjective character (which I construe in terms of “reflexivity” or self-acquaintance)—instead, their subjective character would just derive from their integration into the underlying creature-consciousness, which *ipso facto*

makes the organism to be the subject-pole of the episode ([Schlicht this collection](#), pp. 5-8).

Part of claim (3), as stated, (the part that begins with “since”) is my interpretation of Schlicht, since he does not spell out the claim in great detail, given limitations of space. So my arguments directed at that interpretation may not target exactly what Schlicht had in mind. But claims (1) and (2) are stated very clearly ([Schlicht this collection](#), pp. 5-6, 8-9). In this reply, I will take issue with these three claims and discuss some of Schlicht’s other claims in relation to them.

2 The phenomenological subject and the organism

I readily admit (and did so in the contribution) that the claim that the subject of consciousness is the episode (or stream) of consciousness itself is rather counterintuitive.¹ However, part of this counterintuitiveness can be ameliorated easily enough. To begin with, I should have made clearer that I was talking about what I like to call the “phenomenological subject” of consciousness. I did use the phrase, but I did not explain it and should have done so. The phenomenological subject is just *that to which* the objects of phenomenal consciousness *seem* to appear. In other words, granted that consciousness seems to have a subject-object, relational structure, the phenomenological subject is just the subject-pole of conscious experience in so far as it is given (reflectively as well as pre-reflectively).

Now, suppose we interpreted the Humean “no-self” intuition in the strongest possible way. In that case, we would conclude that there is no phenomenological subject at all. As I argued in the paper, I think the Humean intuition is not to be dismissed. However, I do think that the subject-object polarity of consciousness is a datum and not projected or inferred. If one is already sympathetic with the idea that consciousness is always aware of itself (is its own “secondary object”, as Brentano put it (see

[Brentano 1995](#), pp. 128ff.), or is always non-positionally conscious of itself, as Sartre put it (see [Sartre 2004](#), p. 8)), then it is not much of a stretch to identify this feature of “reflexivity” with subjective character and, if we must reify, make the episode or stream into the phenomenological subject. As long as one understands that by “subject” here, I just mean the subject-pole of conscious experience, a pole that one is phenomenally conscious of, then one will get my meaning. Given a commitment both to some version of the Humean intuition and to the intuition that one is phenomenally aware that consciousness has a subject-object polarity, one will need to resolve the tension between these in some way. Self-representationalism and self-acquaintance theories can do this in a very elegant way, I argued, a way that neither first-order nor higher-order representationalisms can. All of this is compatible with the evident fact that we normally experience ourselves as having a body in space that bears various relations to objects in space; but the phenomenological subject of consciousness should not, in my view, be identified with the body or with a representation of the body.

One might, of course, use “subject of consciousness” in different ways. One might, for example, mean “that organism or system to which we attribute consciousness” or “that which is the substrate of consciousness in an organism”. We might then speak of the “metaphysical subject” or “ontological subject” of consciousness, rather than the “phenomenological subject”. The metaphysical subject of consciousness need not appear to or be represented in consciousness. I have nothing against the idea that the organism (or a set of sub-processes of it) is the metaphysical subject of consciousness. I grant, moreover, that we normally speak in such a way that the grammatical subject of attributions of consciousness (or conscious mental states) is a noun that refers to an organism. We say things like “Skipper, my dog, sees his food coming” or “The bird saw me walking toward it and became frightened”. Indeed, insofar as consciousness is a property of or process going on in the brain of organism, there is nothing erroneous about such attributions. However, given the

¹ Indeed, after hearing me present a version of the target paper (at TCU in 2014), Michael Tye told me that even on a charitable reading, the claim lacks a truth value. At least, that’s what I understood him to mean.

falsity of animism and the commitment to what I like to call encephalism (the view that consciousness resides in the brain), the ontology of consciousness cannot just be read off of the grammar of such attributions, not that Schlicht is suggesting that it could.

It would, however, be more accurate to say in the Skipper case that there is a process of phenomenal visual consciousness having such-and-such representational content and being connected in such-and-such a way with Skipper's volitional, appetitive, and motor systems (all, of course, related to Skipper's organismic homeostatic systems) going on in Skipper's brain when, in normal conditions, the food is presented to him. Of course, I just referred to Skipper *qua* organism multiple times in reformulating this apparently simple attribution, but that just has the effect of roughly localizing the conscious process and, of course, connecting it to Skipper's behaviors and functions as an organism. No one should deny that consciousness, as it has arisen in organisms with an evolutionary origin, has a biological function, though it is highly debatable that consciousness should be defined or analyzed in terms of such functions. It may well be that it serves these organismic functions but could exist in substrates that do not have them or need them. In fact, I would put my money on the claim that artificial consciousness is possible in systems whose homeostatic functions can be carried out in a way that its consciousness does not contribute to *at all*. But that is a debate for another time.²

The counterintuitiveness of claiming that the phenomenological subject of consciousness is the episode or stream of consciousness might derive in part from the oddness such a view would seem to introduce into our quotidian attributions of conscious mental states, if we were to try to make our ways of speaking match this theory. It would be rather odd indeed to say, "Skipper's current episode of consciousness sees the food coming". But the view I defend does not really legitimate such locutions. Those attributions run together the phenomenological and

metaphysical notions of "the subject". The sense of counterintuitiveness that comes from saying "the episode sees..." (etc.) stems from the fact that the episode is *not* the metaphysical subject.

When we make normal attributions of conscious mental states to a creature, we encode information about the location or individuation of the conscious episode (and this gets construed as "ownership"—it's Skipper's seeing), information about the representational content and modality or attitude the episode involves (food and seeing, respectively, in this case), and a sort of folk theory about the relational structure of consciousness. That folk theory puts whole organisms or agents, as it were, "behind" the conscious mental state, as the point of view or subject pole from which the experience emanates or, to use another metaphor famously attacked by Dennett (see 1991, ch. 5), as the spectator in the "Cartesian Theater". That folk theory is hopelessly homuncular, it seems to me. It offers no analysis of what a subject is and gives no hint as to what the real conditions of unity are for either organisms or subjects. (By contrast, self-representational and self-acquaintance theories try to preserve what is right about the Cartesian Theater intuition while avoiding a commitment to homuncularism.)

When we say "Skipper sees" we do not really mean that Skipper is, as it were, behind some sort of internal telescope looking out of his eyes or at some internal screen, though the first theory of seeing that some kids come up with is indeed the homuncular and regressive one according to which there is a little person in our heads looking at just such an internal screen. It seems that what is encoded in the folk theory implicit in our normal conscious mental state attributions is something like a homuncular projection of our third-person experience of other conscious organisms onto the organism's first-person experience. In other words, we see Skipper with his excited behavior and the food out there in front of him, some distance from his body; we then imagine that this relationship that we see "sideways on" (to borrow a phrase from John McDowell, see e.g., McDowell 1994, pp. 34-36) is, so to speak, rotated 90 degrees

² It is also possible that something other than consciousness could carry out most (but not all) of the functions consciousness performs in us and other organisms; this is also a debate for another time.

and moved inside Skipper's head—with Skipper as an irreducible agent assuming the position behind the eyepiece of his internal periscope.

That may be a bit too fanciful an exercise in the conceptual archaeology of folk-psychological mental state ascriptions, but the main point is just that our normal attributions of conscious mentality seem to run together generally accurate information about individuation or location ("ownership"), content, and attitude with a naïve and homuncular picture of subjective character. So, yes, it is true, I would say, that the organism is the subject of consciousness in the sense that conscious episodes (so far anyway) take place in organisms (actually in their brains) and causally depend upon organismic metabolic and homeostatic processes for their existence. In this sense, the organism is the *metaphysical* subject of consciousness, and this is properly reflected in our usual mental state attributions. However, I would emphatically (perhaps even hysterically) deny that the whole organism could be the *phenomenological* subject of consciousness. This is something also reflected in our usual attributions, but this is because the metaphysical and phenomenological subjects are simply conflated by folk psychology. The whole organism could not be the phenomenological subject for two reasons.

First, if one agrees with me, as Schlicht seems to, that subjective character and the subject-object polarity are phenomenally manifest even in pre-reflective consciousness, but one adds to this the claim that the phenomenological subject is the organism, then it would seem to follow that we are always aware of ourselves *qua* organism when we are consciously aware of anything—since, again, all consciousness by hypothesis has subjective character. Now, this could mean either that we represent ourselves *qua* organism or that we are acquainted with ourselves (and we are, in fact, organisms). Surely we do not, at the level of consciousness, represent ourselves *qua* organisms all the time, unless all one means by that is that consciousness has a biological function of some sort (that is, in, say, the teleofunctional sense, consciousness is "about" the organism and its ongoing relationship to the world). The latter is

undoubtedly true, but that is not a phenomenological characterization of subjective character; rather it is a thesis about the function of consciousness and its relation to organismal homeostasis. Of course, one could make an identity claim according to which subjective character (as experienced) really just is the suitably integrated representation of the organism, but this then would mean that one is embracing some form of P-theory (a theory according to which conscious representational states are distinct from non-conscious ones in part because they target some privileged object, e.g., the organism, a substantial self). If the claim is taken to mean that the organism is self-acquainted, then I might be willing to agree depending on the spin one puts on that claim.

One might just mean that there is some sub-process of the organism that is self-acquainted (that is self-manifesting or directly phenomenally self-representing, if one prefers). If that is all that is meant by the claim, then I can agree. Something like this is exactly the position I defend in the paper. After all, the central claim was just that consciousness is self-acquainted. And it was an unstated assumption of the paper that consciousness is a sub-process of the brain, and the brain a part of the organism. If, on the other hand, one means that the *whole* organism is directly acquainted with itself, this seems to me to be either an unexamined endorsement of commonsense, homuncular ways of making conscious mental state attributions (criticized above) or the claim that consciousness *necessarily* involves the entire organism.

The latter disjunct seems as false to me as the former. Yes, the prolonged existence of consciousness depends on the prolonged operation of the essential metabolic and homeostatic functions. And, indeed, almost certainly the metabolic functions that support synaptic transmission and some other basic neuronal processes are *sine qua non* for consciousness as it happens to be implemented in human and animal brains. But none of the specific means whereby our bodies support these functions have to be in place, it seems to me.

We can have artificial hearts and artificial respiration. In principle, we could offload all the

metabolic processes outside those internal to the nervous system to non-natural machines. And we could even, in principle, replace the natural generation of essential neurotransmitters with their artificial synthesis and, possibly, artificial projections for distributing them in the brain properly. Homeostasis could be maintained artificially and, in principle, without the relevant brainstem nuclei needing to do anything anymore (unless, of course, some of their activities just *happen*, for totally contingent evolutionary reasons, to be constitutively necessary for the occurrence of consciousness in brains like ours).

In short, it is certainly physically possible (though technologically beyond our current means) to keep a brain alive and operating in a “vat”! As long as we maintain those processes that are the neural correlates of (are identical to, in my view) consciousness, there would be consciousness in that brain in that vat. I am sorry, but all the evidence indicates that encephalism is true. And it seems to me to be a sort of externalist fetishism to think that consciousness literally extends beyond the brain (save by intentionality and causal interfacing). As Dan Lloyd says, we *already are* brains in vats! The cranium is the vat! (See [Lloyd 2004](#) pp. 244-245.)

Haven’t we learned from dreams, hallucinations, ALS, direct brain stimulation, and locked-in syndrome that consciousness does not need anything but the relevant brain functions to exist? Of course, the functioning brain depends upon a properly functioning body, but this does not mean that consciousness should be identified with those (other) bodily functions in some way. If we go this route, adding bodily correlates to neural correlates, when the latter causally and distally depend on the former, what is stop us from adding everything the body depends upon to our list of correlates (the gravitational constant, the bonding properties of molecules, the stability of the proton, etc.)?

True, at a certain level of analysis, it is hard to say precisely where the nervous system ends and the rest of the body begins. But then the same can be said for the body and the rest of the world (especially given that we routinely

appropriate, by breathing and eating, parts of that world). The boundaries are fuzzy at a certain scale, but this does not mean we should say there are no boundaries at all. Moreover, causal dependencies and interdependencies are myriad, but the causal relation is not the parthood relation—we cannot infer from “X causally depends on Y” that Y is a part of X. Human consciousness, as it is currently implemented, causally depends on respiration, but this does not mean that respiration is part of consciousness or that the physiological correlates of respiration are also correlates of consciousness.

If it is true that consciousness resides in the brain and depends on our specific homeostatic and metabolic processes only contingently (and I mean physically contingently, not merely logically contingently), then saying that consciousness requires an organism as a subject-pole starts to look a little fishier. This brings me to the second main reason why I would reject the claim that the whole organism is the phenomenological subject of consciousness.

What, after all, is an organism? It is clearly not just a collection of parts. We can agree with the Aristotelian tradition that it is a functional unity. We can agree that it is a system that, in virtue of its form or organization, is able to give rise to temporal successors (and I don’t necessarily mean offspring) that maintain that form or organization (at least for a while and at least within the range of conditions it evolved to live in). The matter always changes, but the form remains, from cradle to grave.

The organism takes matter from its environment to keep its processes going. And relative to its function of maintaining homeostasis (thereby giving rise to temporal successors that have the same organization it does) and to the scale at which those functions are, so to say, visible, we can truly say that the organism behaves like a goal-directed *whole* with interdependent parts and processes (and organs—the heart needs the lungs; the lungs need the heart; the kidneys need the stomach, etc.). This is all fine and dandy. But it is clear that organs are not to be identified with the whole organism. Skipper is not Skipper’s heart, though without it (or some suitable artificial replacement) little

Skippy would soon cease to be an organism at all and eventually become soil (or parts of other organisms).

More to point here, the brain is not the organism. Consciousness resides in the brain. We could, in principle, preserve consciousness simply by preserving the functioning brain. If this is true, then you do not need an organism for there to be consciousness, you just need a suitable *organ*—the brain.

I would never deny, of course, that we normally represent ourselves as having a body and relating to a world through that body. No doubt about it. Moreover, I believe it is metaphysically necessary that any consciousness be embodied in some substrate and that this embodiment configures consciousness in a way that is phenomenologically accessible to a certain extent. And, of course, we are organisms with a certain natural history. All of this “facticity” does indeed configure our consciousness to one degree or another. If all the organismal claim comes down to is that conscious beings are necessarily acquainted with their own contingent embodiment in a certain manner, then I will wholeheartedly agree. If it means that, contingently, consciousness evolved out of and is still connected to basic homeostatic functions (in some way), I will regard the claim as a not implausible hypothesis to be investigated.

This last claim is not, it seems to me, exactly what Damasio (see [Damasio 2010](#), ch. 2) and philosophers who follow him on this point, like [Charles Nussbaum \(2003\)](#) and apparently Schlicht himself, mean. They want to say something stronger. Like Francisco Varela (who possibly influenced Damasio on this point, see [Damasio 1999](#), p. 347; cf. [Varela 1979](#); [Maturana & Varela 1980](#)), they seem to want to connect consciousness essentially in the constitutive sense to the kinds of processes that are involved in homeostasis and the very emergence of an internal organism/non-organism distinction. It may well be that the subject/object distinction apparent in consciousness is, in evolutionary terms, some sort of extrapolation of this more basic distinction. It is clear, however, that, whatever the exact relation, the organism/non-organism distinction in, say, the immune sys-

tem, cannot just be identified with the subject/object distinction in consciousness. On Damasio’s view (as I understand it anyway) consciousness arises out of multiple, integrated layers of representation of the organism/non-organism distinction, where this representation itself has a certain regulatory function.

I certainly agree with [Schlicht \(this collection, p. 7\)](#) that, for Damasio, organismal and objectual representations have to be integrated in the right way for there to be something it is like for the organism, and I did not mean to suggest otherwise. This does not, however, make Damasio’s theory a non-representationalist theory (no more than the poise requirement makes Tye’s theory non-representationalist (see [Tye 1995](#), p. 138, [2000](#), p. 62)). As long as representation is considered a necessary condition for consciousness, the theory is representationalist, by my lights. And since the relevant representations, in Damasio’s theory, include, centrally, representations of the organism, it still qualifies as a P-theory in my sense—the organism is a privileged object. The representations that, when integrated, constitute consciousness, must include representations of the organism, according to the theory. Whatever else is represented in conscious mental states, on this sort of view, the organism certainly is. And the organism (ultimately in its guise as the “core self”) could thus serve as the phenomenological subject of consciousness.

I do understand how such a theory attempts to capture subjective character. In effect, it bundles it up into an object of a special sort that is always represented (one way or another) in any conscious mental state. For various reasons (e.g., the Fichte-Shoemaker Regress, see [Henrich 1982](#); [Frank 2004](#); [Frank 2007](#), pp. 157ff.; [Shoemaker 1968](#)), I do not think that such a theory can do the trick. Briefly, it is not enough simply to represent some object that you just happen to be. That is not sufficient to ground the manifest indexicality (the “I am this, here, now” aspect) of our conscious experience. Also, it is somewhat puzzling to require that consciousness representations have to have some specific type of content. Consciousness seems to be so flexible in this regard, that is

odd to think that there is such a “magic” object of representation. By contrast, I view subjective character (“reflexivity”) as a formal or structural feature of consciousness and not as a matter of representing some object or other (whether consciously or unconsciously)—including “the organism”. In fact, I believe such views, while on the right track relative to views that disregard subjective character, get the cart before the horse. *The representation of oneself as a self or an particular organism depends upon reflexivity, and not the other way around.*

For an organism (call it O) to benefit from representing an organism interacting with the world and with other organisms, there must be some way that it encodes that *it* is O (and not anything else). In effect, it needs a “you are here” (or rather “I am here”) dot on its map, a kind of “fixed-point” (see [Ismael 2007](#)). The mapper needs to know where *it* is on the map it has made. On pain of regress, it cannot derive a representation like “I must be here and not there, this organism on the map and not that one” without having some antecedent, unmediated self-reference or primitive self-knowledge (again, see [Shoemaker 1968](#)). Without this direct self-reference, the best we could hope for is a system that just happens to control itself by representing something that resembles itself in the relevant ways. Such a system might as well be controlling an exact duplicate in a duplicate room next door. We do not get manifest indexicality, self-location, or subjective character out of this. A system built up around reflexivity or direct self-reference (or primitive self-knowledge, if you prefer), would have all the control functions of a system that lacks it as well as these other features of subjectivity. Following a suggestion by Metzinger (personal communication), though with a certain modification, I would be happy to call this kind of reflexivity “prepersonal”. As such, it is the basis for one’s conscious representation of oneself as a person, organism, or anything else for that matter; but it is not essentially the representation of a person or an organism. Rather, it is the reflexivity of a process that happens to be housed in an organism (and in an organ within that organism) and that allows that organism to self-locate in a

multiplicity of spaces (physical, social, semantic, etc.).

I would add that that reflexivity could not itself be purely representational, as I argued in the paper. It very well could be, however, that reflexivity was first achieved in the evolution of organismic control systems and that these control systems have everything to do with the maintenance of homeostasis, though this is a contingency. *That* does not seem implausible to me at all. But that is a hypothesis about the evolutionary origin of self-acquaintance, not an account of what self-acquaintance is or how it is routinely generated and supported in the brain.

3 Unity, individuation, and integration

This brings me to Schlicht’s second and third claims. [Schlicht](#) is absolutely right, of course, to press me on the need for an account of the synchronic and diachronic unities of consciousness and for an account of the individuation of episodes and streams of consciousness ([this collection](#), p. 5). I bracketed such worries because I have not worked out any such accounts to my own satisfaction. I do, however, disagree with [Schlicht](#) on the idea that regarding the organism as the subject of consciousness can help with individuation ([this collection](#), p. 6). It is, in my view, not much easier to specify the metaphysical individuation conditions for an organism than it is to do so for an episode of consciousness. And it is problematic to assume that the brain has, so to speak, figured out what these conditions are for us. It is true that the brain must regulate a certain set of functions and processes in order to facilitate the maintenance of homeostasis; and I can even grant that doing so involves “representation” in a teleosemantic or functional-role sense. But this is orthogonal to any issues about the metaphysical individuation of organisms. While it may be easier to say what sort of processes an organism must involve (see above) and to roughly localize those processes than it is to say when one episode of consciousness begins and another ends, it is no easier to provide the *ultimate* metaphysical individuation conditions that ground the identities of the more basic physical processes that

both organisms and consciousness depend upon. At the end of the day we are always left turning our spades with the thought that there just exists a plurality of things in the cosmos—this proton (or bare particular or property or location) is not that one—end of story!

In the case of consciousness there is no *special* problem of metaphysical individuation, if what we are talking about is the fact that these episodes over here are in “my” head (in this brain), and those over there are in “your” head (in that brain). From a purely epistemic point of view, it does indeed seem to be the case that we individuate episodes of consciousness by reference to individuated organisms. But so what? This is a mere contingency. If, say, we saw conscious processes (or their correlates) first, and only with great effort could we locate the organism to which the processes were attached, we would individuate organisms by referring to the streams of consciousness that “own” those organisms. There is what we might call the “epistemic relativity of individuation”. This does not mean that there are no mind-independent facts about what ultimately individuates things. It just means that, since we have no access to what the ultimate individuator is, no particular way we individuate something should be regarded as privileged. We are guided by practical and interest-relative considerations. We might as well talk about the acorns’ squirrel rather than the squirrel’s acorns, but squirrels are more entertaining to us.

In any case, for any physical process, once you drink the metaphysical individuation Kool-Aid, you won’t come back to normal. At some point you’ll just find yourself saying “this is not that”. And anything you scratch (from universes, to stars and planets, to organisms, to molecules and particles) will fall apart in this connection. I believe there are real unities in nature and that conscious mental states are such real unities (it follows that I think certain brain processes are too), but from the epistemic and conceptual point of view Nagarjuna and the Madhyamikas seem to be right—we cannot pinpoint the basic individuator anywhere, they just seem to dissolve upon analysis one way or another (see e.g., Westerhoff 2009). I believe

such individuator exist but that they are inaccessible to us. And, as I argued in the paper, it is a serious confusion to think that in being self-acquainted you *ipso facto* have access to what it is that ultimately individuates you. You are aware of the unity and individuality of your episodes of consciousness, but those features, in turn, could depend on other unities and individuation conditions that you have no access to (cf. Williford 2011, pp. 202-203).

I am as happy as the next bloke to claim that conscious mental episodes arise and eventually give way to the next conscious mental episodes. Perhaps they overlap somehow to form a stream. Or perhaps they are punctate and we could empirically determine their temporal boundaries; I do not know. It is, however, clear that my conscious episode of reading H.P. Lovecraft for nostalgia’s sake before bed is not the conscious episode of eating yogurt for breakfast that I began the day with. But is it the same *organism* that eats in the morning and reads in the evening? It is indeed, given our normal (possibly partly pragmatic) individuation criteria for organisms. One would not have to be Heraclitus, however, to notice that the being that started the day is *quite different* from the one that unwisely decided to read Lovecraft before bed. And at a certain very fine-grained level of analysis, it is a *radically* different being.

We say things like “I ate yogurt this morning” and “I read Lovecraft at the age of 14” and “I am reading Lovecraft right before bed tonight”. We take this “I” to refer to the same organism and to the same “autobiographical self” to use Damasio’s term or same “ego” in Sartre’s sense of the word. But as Sartre pointed out (2004, pp. 7-9), no such temporally extended and dubitable entity could be entirely present in consciousness (so as to serve as its subjective character). Instead we have only a set of processes that remain more or less constant (and of course a causal-historical chain that is not broken). But the whole causal-historical chain does not exist at the present moment. It cannot be packed into a single conscious episode (though it could, of course, be represented in one). Nor can it, as a real pole of identity, exist

throughout all conscious episodes. There is no transcendental ego. Nor is there any transcendental organism. An appeal to the organism does not, just by itself, help us with the diachronic unity problem or the individuation problem. And though the brain may somehow unconsciously always “represent” us as organisms (in perhaps the teleosemantic sense of “represent”), it is evident that we are not always conscious of ourselves as organisms (whatever exactly that is supposed to mean). Yet subjective character is there whenever consciousness is.

Thus, I do not see how the organismal theory helps with either the individuation problem for conscious episodes and streams or the related diachronic unity problem. But it is also not obvious to me how the organismal theory could help us with the synchronic unity issue either—the issue of how different phenomenal and representational contents get integrated into whole, unified conscious mental states or episodes. I fully agree with Schlicht that a brain process (e.g., an “unconscious perception” of something) can be integrated into consciousness, making a new whole, and then possibly slip out again. And I too like Edelman’s and Tononi’s “Dynamic Core” idea (see [Edelman & Tononi 2000](#), Part IV) as a way of conceptualizing this integration and dis-integration. And I would emphatically reject the “micro-consciousness” idea of Zeki (see [Zeki 2007](#)). It does not seem to help to imagine many consciousnesses in the brain that somehow meld together to make a bigger one. (At least I hope it is not like that!)

In connection with this, I, like most people, believe that normally there is only one stream of consciousness existing in a given brain (with split-brain cases perhaps being an exception). I prefer the idea that consciousness is a type of process that has certain generic, essential structural features (temporality, subjective character, phenomenal character, representational character) and certain variable features (this comes down to variability of phenomenal and representational characters at different times). Due to some constantly fluctuating integration process, the phenomenal and representational characters of consciousness are al-

ways in flux, while temporality and subjective character remain invariant. Moreover, phenomenal and representational characters are such that they can, so to speak, expand and contract.

I can be hearing Bach’s *Musikalisches Opfer* while staring at a Jackson Pollock painting. I can then close my eyes so that only the beauty remains in my consciousness. If I open them again, then the visual horror will be reintegrated into it. When I closed my eyes there was “contraction”; when I opened them again, “expansion” back to the “size” the experience was before I closed them. There must indeed be something that accounts for this integration process and the resulting synchronic, differentiated unity of consciousness.

I completely agree with Schlicht that we need a theory that allows us to understand how something enters consciousness and how it gets integrated into a whole with other things that have already entered consciousness, but I do not know what that theory. Moreover, I do not know if it is possible to have a conscious experience of but one sensory quality in one modality (say, the auditory consciousness of a pure C tone) without any other thoughts, imaginings, perceptions, or anything else. I would say, though, that if one could, that episode of consciousness would still have subjective character (and temporality).

Though I agree that we all need an integration story, it is quite unclear to me that that story alone will give us a story about subjective character, unless, like [Van Gulick](#) (see e.g., [2004](#); cf. [Metzinger 1995](#)), one thinks that subjective character as a kind of reflexivity emerges out of integration. That may be, though I have never understood how, exactly, that is supposed to happen on Van Gulick’s view (though I do have some sense of what he means). It is not clear to me how this is supposed to happen on [Schlicht’s](#) view either (see [this collection](#), p. 11). But I would be quite pleased if an account like this could be made to work, since deriving reflexivity from integration would, it seems to me, be a theoretical advance.

In any case, for me, subjective character as reflexivity is a phenomenological given. It is

part of the data set from which I begin. It is my “Phenomenological Muse”. I could be wrong about its importance, of course, but until I am shown that, I will explore the model space that is appropriate to that intuition and leave it to empirical testing to determine whether or not the Muse was lying to me. (I will add, in a purely psychologistic vein, that many philosophers are allergic to reflexivity for purely irrational reasons. They just find it odd or too complicated or too puzzling. So they see it as an advantage if they can offer an account that gets around the need for it. For me this is like taking Marlon Brando out of *Apocalypse Now* or preferring decaf coffee to the real deal.)

As I see it, subjective character is like a universal or form. It is just the “reflexive” structure of consciousness. It is not to be reified into an entity (or refried like a bean, for that matter). If we think of it as an entity, we will find ourselves puzzling over questions about momentary subjects and how all these different subjects relate to each other over time and at a time. This is a confusion in my view. Yes, our normal use of language and the naïve ontology it encodes demand entities and substances to correspond to our nouns! But consciousness is not an entity or substance. It is a process; it is more like a wave than the medium the wave requires. It has a certain structure and a certain dynamical profile. Subjective character is, like temporality, an ever-replicated form that, in my view, is necessary for all consciousness. There is no subject *entity* strictly speaking. When there is an episode of looking back down the tunnel of previous conscious episodes that are connected in the normal way to that very episode of conscious looking, individuated subjective character is always seen. Just in terms of subjective character, all the episodes are qualitatively identical. This helps reinforce the illusion of a stable, continuous subject entity.³ Again, there is no such entity. There is just a common form or structure living in the many different tokens. After Parfit (1984) and the Buddhists, we might say that this at once helps dissolve the thing we

once thought so substantial and important and draws us closer to other tokens, no matter what stream they happen to be in.

Finally, it seems to me that Schlicht (this collection, p. 7) must take “creature-consciousness” to be more fundamental than “state-consciousness” (or “episode-consciousness”), whereas I would adhere to the usual idea that phenomenally conscious creatures are just those that host episodes (or states) of consciousness. I don’t see how reversing this order helps.

4 Conclusion

To recapitulate: (1) I do not see what is gained, either in relation to the individuation problem or the unity and integration problems, by regarding the organism as the phenomenological subject of consciousness. (2) I understand how P-theories attempt to do this by making the organism part of the representational content of every episode of consciousness, but I do not find those theories plausible or helpful, even if we stress the integration aspect of the theories (which does not make them cease to be P-theories). (3) I was less clear on how Schlicht thinks that an integration theory could account for subjective character if it deviates from the Damasio-style theory or from Van Gulick’s HOGS model, which latter I have also always found a little hard to understand, though I am in sympathy with it. (4) I would emphatically deny the existence of Zeki-style micro-consciousnesses; rather, I believe there is (normally) only one stream of consciousness per brain—and that stream can “expand and contract” as more or less gets integrated into it. (5) We do need an account of how unconscious processes get integrated into consciousness and of both diachronic and synchronic unity; but I am not prepared to offer such an account at present. (6) Regardless of how such an account goes, I take reflexivity (self-acquaintance) to be an essential structural feature of all consciousness; and I take it to be a phenomenological datum. All streams of consciousness are immediately aware of themselves, and that is the foundation of all other forms of self-representation, autobiographical cognition, and so on. (7) This reflexivity is subjective

³ See Hofstadter 2007, ch. 7, “The Epi Phenomenon”, for a nice analogy in this connection: the illusion of a marble in the center of a box of envelopes arises just from the stacking of the envelopes (with their repeated structure).

character (for-me-ness), but it is a mistake to turn this structural feature into a kind of entity or homunculus. Thus in saying that the episode is the phenomenological subject, I am offering a non-homuncular account of the subject of consciousness. This ought to reduce a little bit of the weirdness of my claim that the episode is the phenomenological subject. (8) In other senses of “subject”, it is undoubtedly correct to say that the subject of consciousness is the organism, since it is (so far) organisms that have consciousness. However, strictly speaking, consciousness is a sub-process of the organism and lives in one of its organs—the brain. (9) Since we could, in principle, have conscious, functioning brains without the rest of the organism, it seems to follow that the organism is not the phenomenological subject—unless one adopts a P-theory according to which the privileged object we represent is just the organism we happen to be; but see (2) above.

Acknowledgments

I would like to thank Thomas Metzinger and Jennifer Windt and the MIND team, once again, for organizing this volume. Thomas and Jennifer also gave me very valuable feedback on a draft of this reply—thank you both for that! I would also like to thank Tobias Schlicht for writing such a gracious, challenging, and stimulating commentary. I hope he knows that I honestly do not think he is irrationally allergic to reflexivity or that he embraces a fetishistic externalism! Cheers, Tobias!

References

- Brentano, F. (1995). *Psychology from an empirical standpoint* (trans. A.C. Rancurello, D.B. Terrell and L.L. McAlister). London, UK: Routledge.
- Damasio, A. R. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. New York, NY: Harcourt.
- (2010). *Self comes to mind*. New York, NY: Pantheon Books.
- Dennett, D. C. (1991). *Consciousness explained*. Boston, MA: Little.
- Edelman, G. M. & Tononi, G. (2000). *A universe of consciousness: How matter becomes imagination*. New York, NY: Basic books.
- Frank, M. (2004). Fragments of a history of the theory of self-consciousness from Kant to Kierkegaard. *Critical Horizons*, 5 (1), 53-136.
- (2007). Non-objectal subjectivity. *Journal of Consciousness Studies*, 14 (5-6), 152-173.
- Henrich, D. (1982). Fichte’s original insight. In D. Christensen (Ed.) *Contemporary German Philosophy, Vol. 1* (pp. 15-53). University Park, PA: Penn State University Press.
- Hofstadter, D. R. (2007). *I am a strange loop*. New York, NY: Basic Books.
- Ismael, J. (2007). *The situated self*. Oxford, UK: Oxford University Press.
- Lloyd, D. E. (2004). *Radiant cool: A novel theory of consciousness*. Cambridge, MA: MIT Press.
- Maturana, H. R. & Varela, F. J. (1980). *Autopoiesis and cognition: The realization of the living*. Berlin, GER: Springer.
- McDowell, J. (1994). *Mind and world*. Cambridge, MA: Harvard University Press.
- Metzinger, T. (1995). Faster than thought: Holism, homogeneity, and temporal coding. In T. Metzinger (Ed.) *Conscious experience* (pp. 425-461). Thorverton, UK: Imprint Academic.
- Nussbaum, C. (2003). Another look at functionalism and the emotions. *Brain and Mind*, 4 (3), 353-383.
[10.1023/B:BRAM.0000005469.62248.00](https://doi.org/10.1023/B:BRAM.0000005469.62248.00)
- Parfit, D. (1984). *Reasons and persons*. Oxford, UK: Oxford University Press.
- Sartre, J. P. (2004). *The transcendence of the ego: A sketch for a phenomenological description* (trans. Brown, A.). London, UK: Routledge.
- Schlicht, T. (2015). Explaining subjective character: Representation, reflexivity, or integration? In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.

- Shoemaker, S. (1968). Self-reference and self-awareness. *Journal of Philosophy*, 65 (19), 555-567.
- Tye, M. (1995). *Ten problems of consciousness: A representational theory of the phenomenal mind*. Cambridge, MA: MIT Press.
- (2000). *Consciousness, color, and content*. Cambridge, MA: MIT Press.
- Van Gulick, R. (2004). Higher-order global states (HOGS) An alternative higher-order model. In R. Genaro (Ed.) *Higher-order theories of consciousness* (pp. 67-92). Amsterdam, NL: John Benjamins.
- Varela, F. J. (1979). *Principles of biological autonomy*. New York, NY: North Holland.
- Westerhoff, J. (2009). *Nagarjuna's Madhyamaka: A philosophical introduction*. Oxford, UK: Oxford University Press.
- Williford, K. (2011). Pre-reflective self-consciousness and the autobiographical ego. In J. Webber (Ed.) *Reading Sartre* (pp. 195-210). London, UK: Routledge.
- Zeki, S. (2007). A theory of micro-consciousness. In M. Velmans & S. Schneider (Eds.) *The Blackwell companion to consciousness* (pp. 580-588). Oxford, UK: Blackwell.